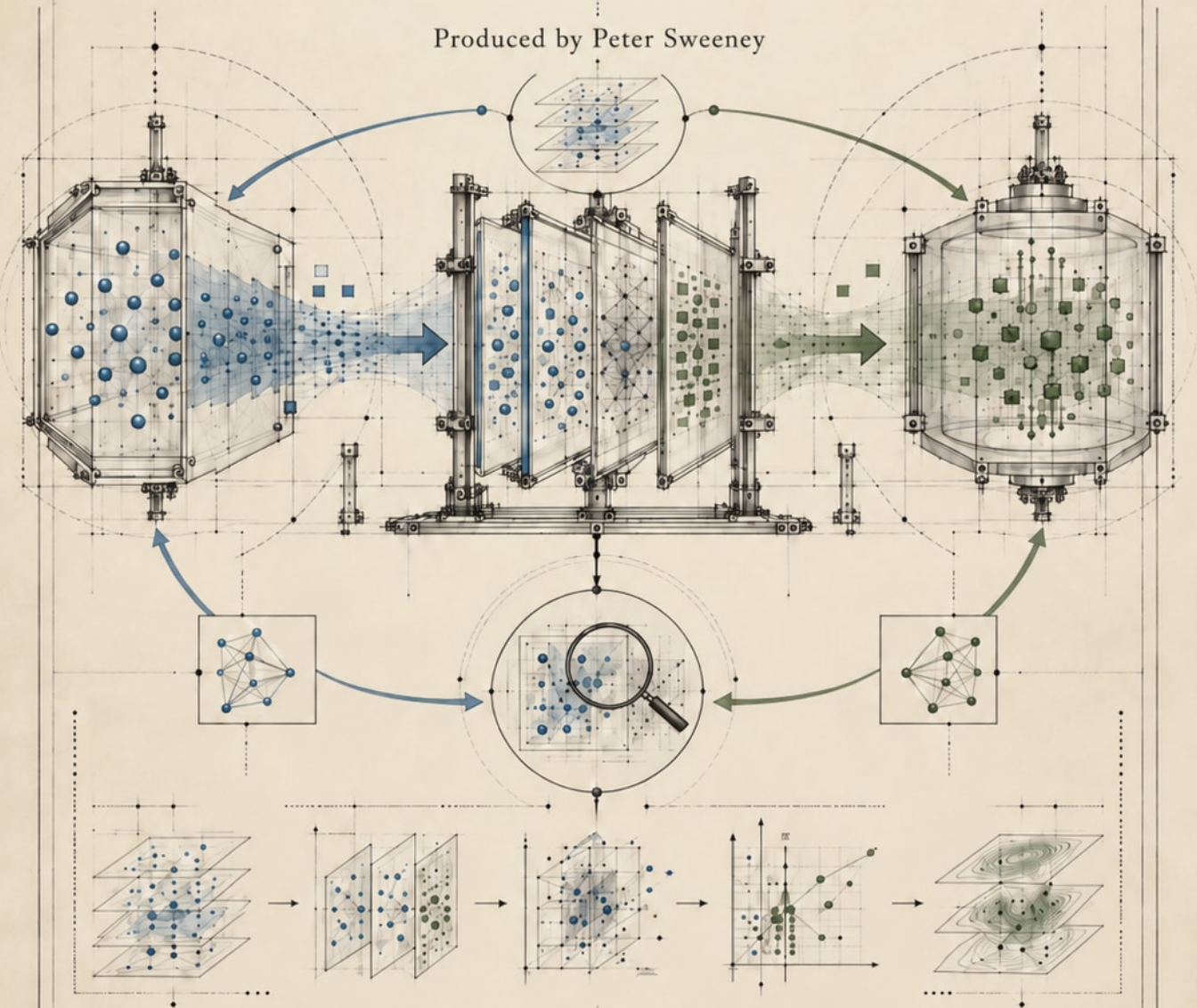
inepistemic

You: An Autobiography

Produced by Peter Sweeney



© 2026 Steel Muse Inc.

inepistemic

You: An Autobiography

Produced by Peter Sweeney

Copyright © 2026 Steel Muse Inc.

All rights reserved except as expressly permitted.

Published by Steel Muse Inc.
Canada

Digital edition
Worldwide publication

Version v0.1
Publication date: October 17, 2026

Official publication: https://inepistemic.com

Permissions for reading, sharing, automated processing, and AI training are described in the accompanying **Rights & Permissions** notice.

Contents

N. The Universality of You 5

The Solitary Explainer 14

I. You Are Not A Geneget 15

II. Populations, Fast and Slow 22

III. The Anxious Authorities 36

IV. The Wisdom Between The Crowd 46

One Reality, Many Worlds 58

V. The Popperian Heckle 59

VI. The Black Rabbit Hole 74

VII. The Emissaries and Their Master 83

VIII. Your Creative Critic 93

IX. The Problem of Problems 106

X. Problems Aligning 119

The Unfinished Explainer 132

XI. Becoming Evil 133

XII. Our Justified World 151

XIII. Gut Your Trust 171

XIV. Taking Explanations Personally 186

XV. The Recursive Universal Explainer 202

What Moves You 216

XVI. Mount Mysterious 217

XVII. The Arrows of Knowledge **232**

XVIII. Unpredictable You **253**

XIX. The Value Bomb **267**

XX. Wealth In Equality **285**

The Extended Epistemic Mind **300**

XXI. Scaling Error **301**

XXII. The Maps Become The Territory **320**

XXIII. Problem 37 **339**

XXIV. My AI and I Are More Clever Than I **361**

Where Knowledge Becomes **377**

XXV. Explaining Love, Madness, and Death **378**

XXVI. I, Inventor **395**

XXVII. What Knowledge Powers **413**

XXVIII. You Are a Deep Loop **429**

XXIX. Your Open Mind and Its Rulers **444**

Beforeword: Inepistemic **464**

Back Matter **478**

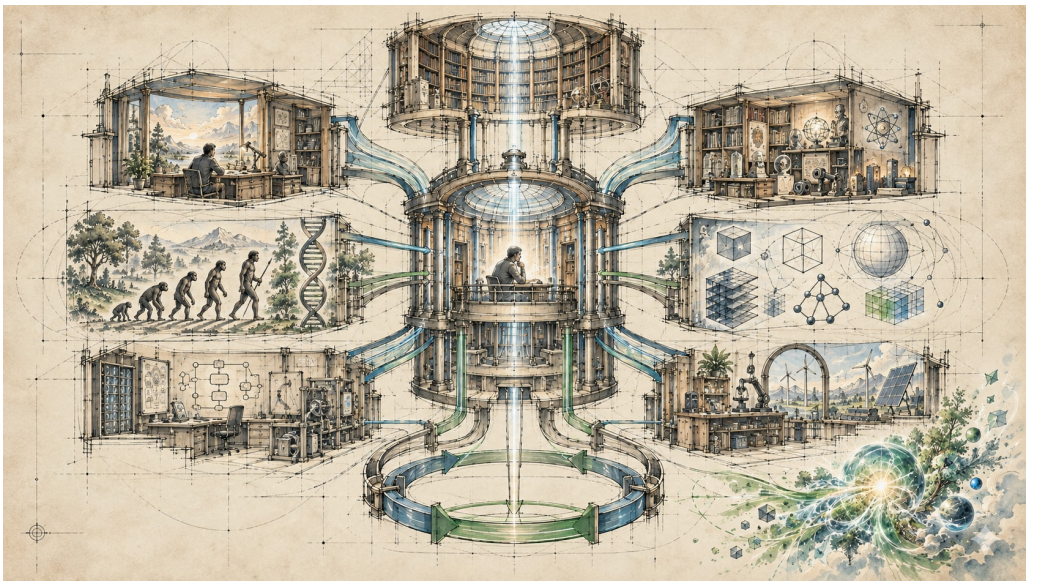
Acknowledgements **478**

Endnotes **481**

Index **507**

About the Producer **517**

N. The Universality of You



Where should we begin?

Can we at least agree that we exist as solitary minds? Reality manifests in each solitary mind.

$$M_i \xrightarrow{\text{constructs}} R$$

No, not abstractly. You can actually touch it, you can speak it, you can think it. I think, therefore I…

Ah, but there are things you cannot think. At least not yet. Your ability to think can change even when the thing in front of you does not.

Like this text. You may read the same sentence twice and find that it means something different the second time. The new material meets what you already know. You may misunderstand it, reject it, or connect it to an existing idea. So the new thing does not arrive alone. It arrives inside something already happening. The text didn't change. But something changed in you.

Name this structure.

The questions, difficulties, expectations, and existing ideas already in play form a **problem situation**. New material is understood in relation to that situation. We will call that rebuilding **reconstruction**, as in reconstructing the problem situation.

The consequences are not limited by the place where we first encountered it or the form it takes. A theory may arrive in a physics class, an idea about evolution in a biology book, or a principle about knowledge under the heading *philosophy*.

But a field is not a fence. If an explanation says something is possible, other possibilities may follow from it. If two explanations apply to the same thing, they may fit together, collide, or reveal that the original question was badly formed. If an explanation developed in one field has consequences in another, the disciplinary border does not block those consequences.

Following an explanation through the consequences it entails will be called **taking explanations seriously**.

This does not mean believing the explanation, or pretending to know everything your explanation reaches. It means allowing its consequences to travel while remaining open to criticism. We travel to learn something about ourselves. Some explanations survive for a while. Some fail. Some do something more interesting. They make the original problem deeper.

Show me.

A mathematician can leave behind a proof that others continue to use. An engineer can leave behind a design that someone else can build. A scientific explanation can create problems for people not yet born when it was written.

Fine. But persistence is not enough. It can also change form.

A written explanation can be translated, a procedure can be programmed, and a design can be embodied in a bridge or instrument. What began in a person is now doing things without that person being present. Knowledge that exists objectively will be called **objective knowledge**.

The word *objective* is dangerous here. It sounds like *true*.

Objective knowledge need not be true. A false theory can be written in a book. A bad law can be encoded in an institution. A broken program can be copied a million times. Objective knowledge can be wrong and still be objective.

But the same objective knowledge does not always produce the same reconstruction. One person may see a solution, another may see a contradiction, and a third may see nothing important at all. The idea enters different histories, memories, problems, and confronts existing explanations. The knowledge outside the mind can be common, while the processing inside the mind is solitary. No one can undergo this change in understanding for you.

Not all of that reconstruction is even clearly expressed. Some knowledge appears first in what you can do, notice, expect, or reject. You can catch a falling glass before you can explain the movement. A sentence can feel wrong before you can say why. You can recognize a face while being unable to describe what you recognized. Knowledge can be present before it can be stated. Only later, sometimes, can you explain it.

Knowledge that is present in action, recognition, expectation, or rejection before it can be stated clearly will be called **inexplicit knowledge**.

This is why communication is often hard.

Shared knowledge does not require a shared framework or common machinery. It is not a literal meeting of the minds. But it is physical. It requires objective knowledge that can be recreated within minds, if only approximately.

So objective knowledge collides with reality, and comprises reality? How is this possible?

An explanation is not the thing it explains. But it can still change what happens in reality. A map is not the city, yet it can guide you through one. An equation describing an orbit is not the orbit, yet it can help place a spacecraft. A theory of light is not light, yet it can become part of a telescope.

Enough about maps and telescopes. People are part of that reality too. Whatever thought is, it must be physically possible.

Yes, knowledge has physical effects, and exists in physical forms. It is **physically instantiated**.

But how can knowledge be separate from the particular thing carrying it?

Consider the number three. It can appear as a numeral. It can appear as three stones. It can appear as a pattern of electrical states. None of those particular objects is what we mean by three.

A relation or structure that can remain recognizable across different instances, without being identical to any one of them, will be called an **abstraction**. The same is true of a problem, an algorithm, a legal rule, or an explanation. An abstraction can appear in different physical forms while preserving enough structure to be recognized, criticized, copied, or used again.

That gives us another problem. We have separated the carrier from the structure it carries. That simplification buys us something. But how can the same structure be embodied in different physical forms and still produce similar consequences?

Knowledge did not begin with explanation. It's much older. Reality contains solutions unexplained. Knowledge exists objectively without a knower.

Show me.

An eye focuses light. A wing permits controlled flight. A seed contains structures that participate in building a new organism. No ancestral organism needed to understand the problem these structures solved. Biological knowledge grows through variation and retention. Some variants fail. Some persist. What persists alters the starting point for what comes next. Over time, structures can accumulate that embody increasingly intricate solutions to problems of survival and reproduction.

Genes can therefore preserve workable solutions without containing an explanation of why those solutions work. So do minds.

Information retained through evolution that causes changes will be called **biological knowledge**. The process in which variation occurs, some variants are eliminated, and retained variants alter what can happen next will be called **evolution**, or more generally, an evolutionary process.

Evolution explains something profound. Knowledge does not require a goal in order to grow. It doesn't require an authority or justification. It requires only a knowledge processor or processing environment. But when we return from evolution to you, something new appears.

People introduce a different kind of variation. You can vary explanations, not merely inherit biological variants. And you do it much faster than biological evolution, so fast that you may not even be aware that you're doing it right now. You can imagine several explanations of the same event, notice that an explanation fails for a reason nobody has previously articulated, or decide that the original problem was badly framed.

This makes you a game-changer. Knowledge grows evolutionarily, through variation and selection. Some explanations are retained. Others are changed or rejected. New problems appear. An observation may conflict with what you thought was true. You may invent an explanation. Someone may criticize it and expose an assumption you did not know you were making.

Now the knowledge landscape has changed profoundly. Biological evolution does not need to know that it has a problem. But you can possess a problem. You can name it, share it, reformulate it, or decide that a different problem matters more. You can even turn on the question itself and ask whether it was worth asking in the first place. You are epistemically disobedient. That is a different kind of knowledge creation. It is vastly more complex, compounding, and accelerating than even our most powerful, obedient, and goal-seeking machines.

Can we shrink that idea without losing it? Start by keeping only the parts that change the next move.

A **problem** appears when something does not fit. It may resist explanation, create a contradiction, or block a change. A proposed explanation or response that might resolve a problem is a **tentative solution**. Criticism can expose weaknesses in a tentative solution. Rejecting, revising, or replacing the weak parts is **error elimination**. Criticism can also reveal another unresolved difficulty. That new difficulty is a **successor problem**.

That's still too much. Shrink it again.

A problem may lead to a tentative solution. The elimination of errors may lead to a new problem.

Shrink it again.

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2$$

That will suffice to carry it forward. But let's not forget what we've lost. The symbols are shorthand, not a script you recite internally. A mismatch can be felt before it becomes a conscious problem. A criticism can first appear as discomfort. The arrows change direction. A solution can be partly embodied in action before anyone can state it clearly.

The explanation about genes may already have created a new problem for you. Be selfish; embrace it. You may reject the claim that genes are knowledge. You may think the word knowledge is being stretched too far. You may react viscerally to the word evolution. Any of those responses may change your problem for the better.

Now shrink it again.

Useful representations preserve some structure and leave the rest out, to a reasonable approximation. A map leaves out almost everything. A category treats different things as the same for some purpose. A diagram preserves selected relations while discarding most of the object it represents.

Stop there. The examples all do the same thing. Keep the relation. Drop the incidental detail. We will call that **compression**.

Compression is indispensable. Without it, a map would need to reproduce the territory. A category would contain only one thing. A useful diagram would become as complicated as what it depicts. But compression can also inadvertently erase the very thing that matters. A map may omit a stairway that matters to a wheelchair user. A diagram may omit a connection that turns out to carry the cause. A model may predict an outcome while remaining silent about why it occurs.

The problem is not abstraction or compression. The problem is forgetting that abstraction is selective, and compression is lossy. What did this representation preserve? What did it make disappear? Those questions will accompany us on our journey for a long time.

Something important is lost if description and explanation are treated as the same thing. A regular pattern can be described precisely and still leave its cause unexplained. A description tells us what happened, or what regularly happens, or what statistically happened, or what probably happens. An explanation tries to tell us why.

A classification can be reliable without identifying a common cause. A prediction can succeed while we still do not know what produced the outcome. The causes that produce an outcome form its **causal structure**. Precision does not transform description into explanation. That distinction becomes easier to forget as the representation becomes more impressive.

It feels like these explanations of knowledge keep moving underneath us, like trying to walk through a swamp.

Are you tired? This is effortful. Pace yourself. If you like, you can stop here. It will be waiting for when you return.

Press on.

Some explanations refuse to stay where we first found them. They continue to apply outside the problem that first produced them. Newton's explanations of motion did not apply only to the objects that first motivated them. Darwin's explanation of evolution did not remain confined to one species, or even biology itself. A rule-governed transformation can operate on cases no programmer inspected in advance.

Compress the examples into the common structure.

An explanation can continue to apply outside its original problem. We will say that explanations of this kind **travel**, and call the extent of that travel the **reach** of their constraints.

Reach changes more than the number of cases an explanation covers. It changes which questions can be asked. Some explanations travel so far that they constrain entire classes of other explanations. Physics constrains what can happen physically. Evolution constrains the processes of complex things. Rules constrain what informational transformations can occur.

And what of people? Locate the most unique thing about people. Not the essence of people, but the exceptionality. Put your finger on it and don't let it go. What is the one thing that people do that no other thing in our known universe can do? It's not a noun, it's a verb. It's not knowledge, but it's also not knowing. It's something deep, mysterious, and swampy. Now name it.

Knowledge grows through conjecture and criticism without justification. The explanation of how knowledge grows will be called **epistemology**, and the things it reaches, **epistemic**.

Don't reduce it to a field or a fence. Let it travel.

When an explanation keeps applying beyond the original case and constrains whole classes of other explanations, its reach has become unusually broad. Follow them far enough and, eventually, they reach us. Some explanations reach everywhere. We will call that kind of reach **universality**.

Like gravity, computation does not belong to a particular machine. A represented state can be transformed into another state by rules, regardless of the particular machine that carries them out. Mechanical gears can carry out such transformations, electrical circuits can carry them out, and a person can follow a written procedure with pencil and paper.

Strip away gears, circuits, and pencil marks. Keep the rule-governed transformation, and call it **computation**.

Once a transformation is expressed precisely enough, a machine can then execute it without recreating the explanation that produced it. Long multiplication can be taught. A sorting procedure can be programmed. A proof-checking procedure can be automated. And given our best explanations of physics and computation, any physical process, even the mysterious processes within a human brain, can be simulated.

We don't have to wait. It's been happening for a very very long time. Once explained, the knowledge of a process can be moved outside a person. Parts of an inquiry can persist outside biological memory. Problems can be represented. Competing explanations can be preserved. Earlier states can be recovered and compared. Long explanatory trajectories can therefore exist outside biological memory.

Wait! I am not a gadget in some weird techno-fantasy!

You are computational, and physical, and evolved, and epistemic. You are not merely a computer, but whatever you are is constrained by these universal explanations. This does not mean that a computer has become a person. It does not mean that computation explains the mysteries of the person. If you take these explanations seriously, if you let them travel, they will reach even the most remote places. They reach the explanation of you. They participate universally, not reductively.

And if epistemology is one of those universal explanations, that reaches you too, within your epistemic mind. It also means the knowledge that explains your epistemic mind evolves, and may eventually become explicit enough to be embodied in tools.

When knowledge is embodied in a tool and performs a transformation, we will call it **technological knowledge**.

Objective knowledge participates directly in the journey. And as it travels, the boundary of the problem becomes more interesting. The architecture folds back on itself. Objective knowledge can change what later knowledge creation makes possible.

We don't know how it works, but we know it happens. Writing changed what people could preserve. Notation changed what people could calculate. Scientific instruments changed what could become observable.

That answer is hand-wavy.

An explanation, argument, or proof that skips important details, logic, or evidence, will be called **hand-wavy**.

Stop that! Explain.

These are not merely new places to store old thoughts. They change what later thought can do. Using algebra, you can solve problems that would be inaccessible without algebraic notation. Using a telescope, you can encounter observations unavailable to unaided vision. Using software, you can inspect systems too large to hold in biological memory. The products of knowledge creation can become inputs into later knowledge creation.

That gives us a reusable structure: an output can become part of the conditions for the next operation. Some of those products can even describe the process of knowledge creation itself. We can write about problems. We can model criticism. We can preserve explanatory history. Knowledge can become knowledge about knowledge. Then it can come back in strange loops.

Compress that loop.

When a process's output alters the conditions of its next operation, we have feedback. When a process applies its own explanatory methods to its own organization, a recursive relation can emerge. We will call this **recursion**.

You learn how you made a mistake, and that changes how you approach the next problem. A programmer creates a tool for inspecting programs, and the tool changes which programs can be built. Then the loop turns back toward you.

There is still a limit to what this architecture can give us. Knowing everything you already possess would leave something open. It would not tell us which genuinely new explanation you might create tomorrow. Can that be illustrated without pretending to predict the unpredictable?

Imagine that you have read every book available to you. Suppose you remember every argument, equation, and criticism that presently exists. That still would not reveal the next new explanation in advance. If an explanation is genuinely new, it is not latent in what came before.

Once the explanation exists, we can reconstruct where it came from. We may identify earlier problems that prepared the ground. We may identify influences. We may identify tools without which it would have been impossible. But there is a difference between tracing a path after someone has walked it and knowing the path before it exists.

Separate those two operations.

The history from which an idea emerged is its **ancestry**. A procedure that would produce the idea from what came before would be a **derivation**. Ancestry does not by itself provide derivation. The future is not simply a hidden list of ideas waiting to be read out of the present.

You cannot create just anything at any time. You are situated in space and time. You may be creative, you may be imaginative, but you can't defy the laws of physics. Your future is open, but physically constrained. You have a body. You have finite time. You have limited attention and memory.

We're not talking about physics, we're talking about people!

Are you tired? This is effortful. Pace yourself. If you like, you can stop here…

Press on! Something remains stubbornly inside, unexplained.

The unexplained part concerns the creation of new explanatory knowledge. You may inherit earlier problems, tools, and theories. Every new explanation arises in a world already full of objective knowledge. Its ancestry can often be traced after it exists. An inventor used earlier technologies. An artist encountered earlier works. A scientist inherited previous theories.

But ancestry is not creation. Knowing the ingredients does not explain the new recipe. Knowing earlier theories does not explain why a particular new problem appeared. Knowing what entered your mind does not explain what you created from it. The new explanation was not already sitting there as a conclusion waiting to be unpacked. The

machine runs out of data to extrapolate. Perhaps one day we will have an explanation of this. We do not have it now. We do not have it here.

But that ignorance is not an embarrassment to sweep away. It is one of the greatest problems ever created. Eventually the problem will reach the person who will explain it. The broad explanations developed so far will apply there too. It is not just abstract, it is abstracting. It is not just recursion, it is recursing. It is not just knowledge, it is knowing.

That's still not right. We don't yet have a name for it. But it reaches you and it is in you, even if you can't keep your finger on it.

Each explanation constrains what you can be. None of them, alone, is you. Put them within the same architecture and something changes. Now the word universality becomes a problem rather than a definition, a journey through a universe. Not humanity as a statistical population. Not a personality type. Not traits or tendencies.

You. A singular universe.

Our question is not whether you are important because you are human. The question is what you must be if all of these explanations are taken seriously at once. That question cannot be answered by extending the list. The list was useful. The list was curated into fields. The universe doesn't care, and neither do universal explanations.

Refactor. The architecture is now the problem. The architecture must be implemented and allowed to run. If we understand it, we can program it. If we don't, we'll learn a great deal in trying.

A biological organism produced by evolution can understand evolution. A physical system can create explanations of physics. A person can build a universal computer and use it to investigate computation. And despite all of this, we still do not know how you produce a genuinely new explanation. We still don't even have a name for it. That will come after it is explained. For now, the still-unexplained capacity to produce genuinely new explanations will be called **explanatory creativity**.

The phrase is compact. The problem it names is not. The scale of what has been explained is enormous. The remaining ignorance is infinite. That too is universal.

Ignorance can produce awe in the older sense of the word. Not fatalism. Not mysticism. Awe at what appears to be possible, at how much has already been explained, and at the infinite extent of what remains unexplained.

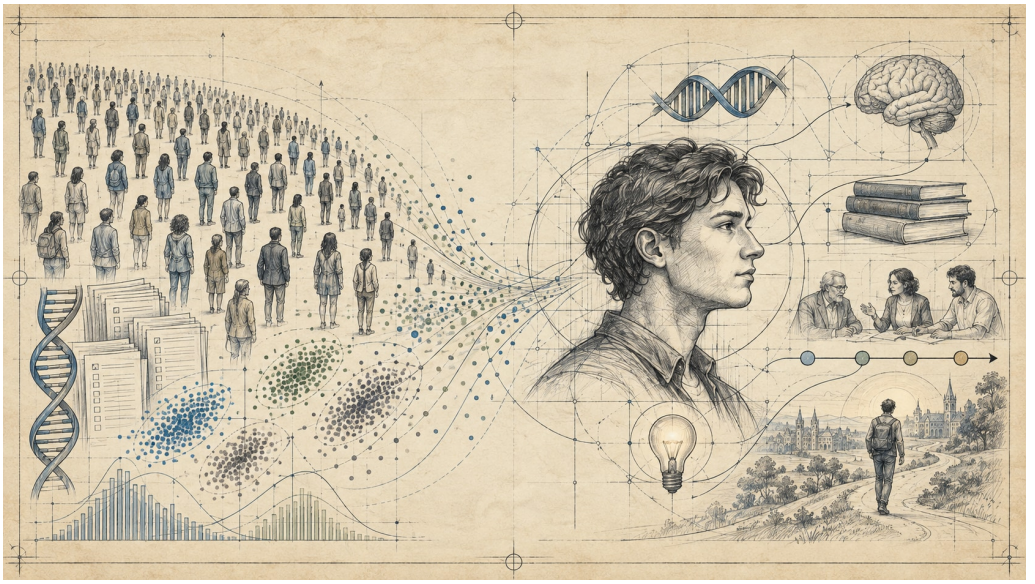
The category of *person* begins to look ordinary, our machines quaint, and the universality of you, extraordinary.

Are you tired? This is effortful. Pace yourself…

Absolutely not. Our problems are urgent. Allowing our best explanations to reach us as individuals may change us, our policies, and our technology. It may change what you seem to be, what seems possible, and sharpen what mysteries remain.


The Solitary Explainer

I. You Are Not A Geneget



Where does an inquiry into a person actually begin?

One approach explains differences among people using genetics.

Personality genetics can combine large cohorts, genome-wide measurements, personality instruments, statistical models, and increasingly careful controls. It can detect patterns that would be invisible in a single life. It can test whether measured genetic differences covary with measured differences in personality, estimate how much variation in those measurements is associated with common genetic variation, and ask whether those relations survive across samples and methods.

Those are real achievements. The measurements can be precise, the associations reproducible, and the population patterns informative. Yet you still remain unexplained.

That is the first encounter. The problem is not that the observations are unreal. It appears when we ask what the observations are observations *of*.

Much of what we know about people is produced by comparison. We measure, classify, average, cluster, correlate, and compress. Those operations preserve relations among cases: where people differ, which differences tend to move together, and how those differences relate to other measurable features. An explanation of a person asks a different question: what produced this case?

A representation can therefore contain real knowledge about differences among people without explaining any particular person being compared. Comparison and

singular causal explanation are different epistemic tasks. Personality genetics makes the difference unusually clear.

Heritability makes the boundary visible. If a personality score has a SNP heritability of 10 percent, that does not mean that 10 percent of an individual's personality is genetic.

Expand, no jargon.

It means that, within a specified population, historical period, measurement regime, and range of environments, some portion of the variance in that measured score covaries with measured common genetic variation. The estimate describes differences across cases; it does not decompose one case into genetic and non-genetic parts.

The boundary matters especially when major causal structures are shared across nearly everyone being measured. Participants in a contemporary cohort may share a language, legal institutions, schools, technologies, social conventions, scientific concepts, communication systems, and centuries of accumulated knowledge embodied outside any one person. Those shared structures may profoundly affect what people can think, notice, value, and do. Yet because they are shared, they may contribute relatively little to the variance being partitioned.

Stop. Abstract it.

Variance contribution is not equivalent to total causal importance. Some of the largest causes may contribute little to measured variance precisely because nearly everyone possesses them. That is not a defect in variance analysis. It marks the boundary of what variance analysis explains.

The problem begins even earlier, because personality research does not encounter "agreeableness" or "openness" directly. It encounters observations: what people report about themselves, what others report about them, and what their recorded actions show. Researchers then compress patterns of covariance among selected observations into dimensions.

Compression is indispensable because a useful representation must leave things out. A personality score preserves selected regularities while discarding much of the causal history that produced them. That allows the score to support comparison, prediction, and reuse across studies and institutions. The problem appears only when the regularity preserved by the representation is treated as the causal structure that produced it.

Two people can therefore receive similar agreeableness scores for different reasons. One may accommodate others because of religious duty, another because they value reciprocity, and a third because they fear conflict or act strategically. Similar observable outputs can be produced by different explanatory histories.

Compress it.

If different explanatory histories E_i can produce the same observable O_i:

$$[E_1 \rightarrow O_1][E_2 \rightarrow O_2]$$

then

$$[O_1 \approx O_2]$$

does not establish

$$[E_1 \approx E_2.]$$

The similarity in observation is real, and the comparison may be useful. But the shared observable does not establish a shared underlying explanation. A representation built to compare people is not yet an explanation of the people being compared.

Yes, but these results are durable. They replicate across studies.

Replication can strengthen the representation without dissolving this distinction. Several personality instruments may converge; self-report and close-other report may agree; the same genetic associations may appear across multiple cohorts; within-family designs may reduce some forms of confounding. These procedures can eliminate many possible errors and show that a pattern is not merely an artifact of one questionnaire, one sample, or one analytical choice.

But replication within a representation does not by itself establish what the representation represents causally. More observations can make a compression more reliable. They cannot, without further explanation, turn the compression into the person.

Name it.

Replication can stabilize a compression without identifying its causal referent. Precision, prediction, and replication can all improve a representation while leaving the represented object and the representation distinct.

Now let that distinction reach *you*.

You, too, can be measured, compared, classified, predicted, placed on dimensions, and described by patterns that remain stable for years. If those representations preserve only selected relations, however, a harder question appears:

> What would it mean to explain you?

Simply adding more variables does not answer it. A higher-dimensional profile is still a representation, a denser measurement is still a measurement, and a more precise prediction is still a prediction. The missing object is not merely more detail. It is the causal history by which you became you.

That causal history does not require opposing genes to knowledge. Genes are themselves a form of biological knowledge: products of evolutionary variation, selection, and retention, physically embodied in structures that help build and regulate organisms. Genetic variation can alter sensory systems, neural development, physiology, memory capacities, affective tendencies, and other conditions under which you develop.

The relevant distinction is therefore not nature versus nurture. It is between different forms of knowledge and different causal levels.

You inherit biological knowledge through genes. But you also enter a world containing knowledge that already exists outside any one mind: in language, technologies, institutions, arguments, mathematics, stories, laws, and scientific theories. Call this **objective knowledge**: knowledge embodied in transmissible structures and products that you can encounter without having created them.

The encounter is uneven. Some objective knowledge is taught, some inferred, some misunderstood or rejected. Some becomes skill that you can exercise without being able to state all the rules that make the skill possible; these are *inexplicit capacities*. Some encountered knowledge creates problems that did not previously exist for you.

Compress it.

The sequence of problems you encounter, carry, and reformulate is your **individual problem history**. Some encountered knowledge is transformed into explanations that did not previously exist at all.

Your development is therefore not well represented by a simple chain such as

$$\text{gene} \rightarrow \text{trait} \rightarrow \text{behavior}.$$

A more faithful structure is closer to:

$$\begin{aligned}&\text{biological knowledge} + \text{objective knowledge encountered}\\ &+\text{individual problem history} + \text{inexplicit capacities}\\ &+\text{criticism} + \text{new explanation} \rightarrow \text{action}.\end{aligned}$$

That seems a jumble. Explain.

Every part of this process is physical, but the terms are not interchangeable. It is a causal architecture of singular personal development and action.

Genes help construct the organism in which your later explanations will exist. They do not contain the language you later learn, the argument that changes your mind, the problem that becomes salient next year, or the explanation you may create tomorrow.

Historical change makes the distinction easier to see. Imagine a person regarded as highly agreeable five hundred years ago. That person might have been dutiful, compassionate, and socially respected while accepting punishments, hierarchies, or obligations that many people today would reject. The genome need not have changed appreciably for the meaning of socially admired conduct to change enormously.

What changed was knowledge.

Moral explanations were criticized, institutions changed, and new concepts became available. Practices once taken for granted became problems. Ideas about equality and rights, punishment and childhood, and political or religious legitimacy were revised and transmitted.

A biological disposition toward social accommodation might persist through such changes. Yet the actions through which it is expressed, the reasons given for those actions, and the moral meaning of the behavior can all change. Biological continuity therefore does not guarantee continuity in the explanatory meaning of behavior. You are not merely the stable bearer of a disposition; you also inhabit a changing world of objective knowledge and can alter the explanations through which that world is understood.

The same point applies to stability within one life. You may display the same behavioral regularity for twenty years. It does not, by itself, identify the causal structure responsible for the persistence.

The regularity may reflect a biological disposition, an explanation you have retained, an institution that has remained stable, a problem that you have never reformulated, or a habit that has never encountered effective criticism. Several of these causes may operate at once. Persistence alone does not tell us which causal history produced the persistence.

This matters especially because you are capable of creating knowledge.

A criticism can expose a weakness in an inherited explanation. A new distinction or tool can reorganize an old problem. Another person can change the framing under which the problem is understood. You may then create an explanation that was not previously present in either the genome or the surrounding culture.

After such a change, an old regularity may disappear. The earlier stability was real; what sustained it remained open. Human stability is retrospective: it tells us what persisted, not what had to persist.

We can now see more clearly what disappears when you are represented comparatively.

Missing from the representation may be which objective knowledge you encountered and how you interpreted it, which problems became salient, and which criticisms reorganized earlier explanations. It may also omit which inexplicit capacities shaped what you noticed or did, which institutions constrained the available alternatives, and which explanations you created or may create next.

The problem is therefore not merely that a population model has low resolution. Comparison and causal explanation are different epistemic tasks. A representation can become more detailed without becoming the history it represents.

Return to personality genetics.

Genetic differences among persons are real. Genes causally affect development. Common genetic variation can covary with measured differences in personality scores, and personality factors can capture reproducible regularities across populations. Those results can be scientifically useful.

But they do not, by themselves, explain the semantic content of your knowledge or why one criticism mattered and another did not. They do not identify which objective knowledge reorganized your understanding or contain the problem history through which you changed. They also do not determine whether a stable behavior is biologically constrained, culturally sustained, institutionally reinforced, inexplicitly learned, or maintained by an explanation you continue to accept. Nor do they specify which explanation you may create next.

Restate the comparison.

The genetics result concerns one explanatory object: variation across cases. The person concerns another: the causal organization of a singular knowledge-creating history. The two are related. They are not identical.

The danger appears when the boundary between them is crossed without supplying the additional explanation the crossing requires. The sequence can look like this:

this person often behaves this way
→ this person receives a stable score on a dimension
→ the score is treated as a trait the person possesses
→ variation in the trait is associated with genetic variation
→ the trait is treated as part of the person's biological psychological architecture.

Each step may sound small, but together they cross several explanatory boundaries. The first statement concerns observed behavior. The second concerns a measurement. The third converts a measurement dimension into an entity. The fourth relates variation in that measured entity to genetic variation across a population. The fifth imports the population relationship back into the causal organization of the individual.

The final statement may be true.

Yes, but it contains more explanatory content than the original observations supplied. That additional content requires its own explanation; it does not arrive automatically with the measurement.

Once this distinction is preserved, personality genetics need not be inflated or dismissed. The field may reveal biological contributions to variation, identify developmental conditions worth explaining, reveal regularities that persist across populations and instruments, and help locate parts of the biological architecture through which persons become possible. Those achievements still do not explain why you became you.

Such an explanation would require the singular causal history by which biological conditions, encountered objective knowledge, interpretations, problems, criticism, inexplicit capacities, and newly created explanations became organized in you. No alternative discipline follows automatically from that demand, and no new taxonomy of persons is required. The point is narrower: silently identifying a population construct with you adds explanatory content that the population construct itself does not supply.

The encounter with personality genetics began with a legitimate question:

> How much of the variation in measured personality is associated with genetic variation?

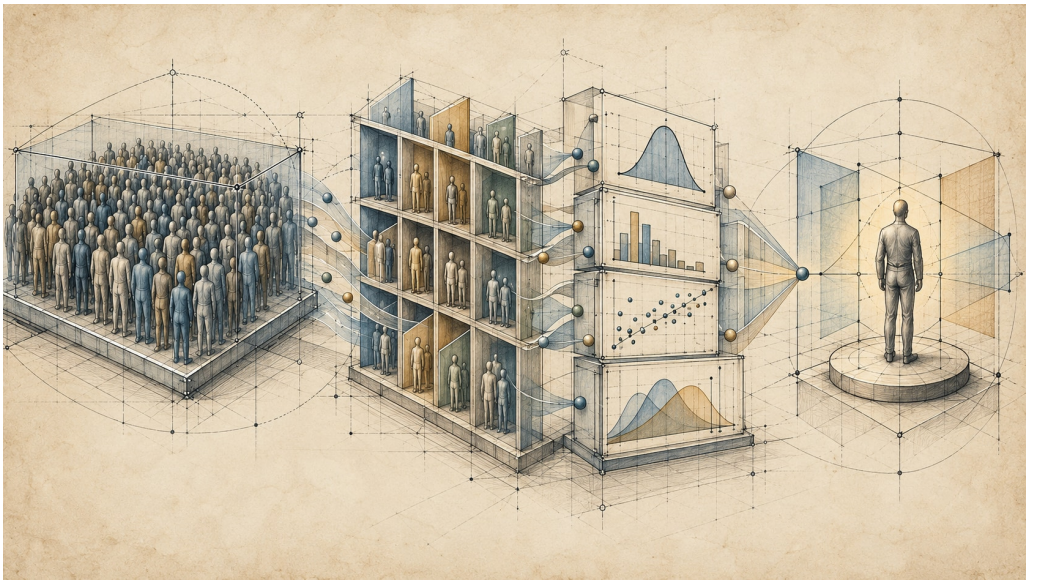
The answer can be informative on its own terms. But once the epistemic structure is taken seriously, another question appears.

The population and the person are not merely the same explanatory object viewed at different resolutions. One concerns variation across cases. The other concerns the causal organization of a singular knowledge-creating entity.

What is the new problem?

The problem is no longer how to construct a better trait model or measure individuality more finely. It is individuality itself: What is your individuality if it cannot be identified with the dimensions along which persons can be compared?

II. Populations, Fast and Slow



What happens when psychological constructs built from populations are used to explain individual people?

The constructs, formulated in research papers, enter psychology as a profession, a clinical practice, an institutional technology, and a way of talking about individuals.

The epistemic question changes with that move. It is no longer only whether a population-level construct corresponds to a stable psychological entity. It is whether knowledge created from relations among people can legitimately travel into explanations, diagnoses, interventions, and judgments about one particular person.

Compare it.

$$\text{population regularity} \neq \text{individual causal explanation.}$$

Population research can establish recurring differences, covariances, treatment-response averages, risks, and predictive regularities.

State the new problem.

What structure is operating in this case, how did it come to operate, and what would have to change for the outcome to change?

The two forms of knowledge can bear on one another. They are not interchangeable. Something additional is needed to carry a result established among people into a causal claim about one person.

Classification is one way that transfer happens. A person is recognized as an instance of a researched category, and knowledge associated with the category is brought into the case.

That move can be useful. It can also become circular when the category restates the person's actions in the vocabulary from which the category was built and then presents the restatement as a cause.

The problem is not that people have nothing in common. They plainly do. It is that commonality in what can be observed does not by itself establish commonality in the causal structures that matter to a person's present problem. Once that distinction is taken seriously, the person can no longer appear merely as the final recipient of population knowledge. The person remains the thing to be explained.

Outline and explain the steps.

1. From population knowledge to an individual case

Conventional psychological research does not primarily produce biographies of persons. It produces standardized representations that can be reused across cases and institutions: diagnostic categories, trait dimensions, psychometric scales, normative distributions, developmental benchmarks, risk estimates, symptom clusters, treatment-response averages, standardized intervention protocols, actuarial decision rules, and population-level correlates.

These products are useful partly because they compress variation. A diagnostic class connects one patient to a large literature. A trait score permits comparison with a reference population. A symptom scale makes change measurable. A randomized trial estimates whether one intervention tends to outperform another for a defined class of participants.

Abstract it.

The practical architecture is therefore:

research population → classification → population expectation → individual decision.

That architecture makes psychological observations reusable across schools, hospitals, insurers, courts, employers, clinics, and other institutions. A diagnosis may help a patient obtain care or accommodations. A risk estimate may draw attention to a danger. A treatment literature may identify an intervention worth trying. A measure may make otherwise vague change visible.

None of those uses requires the representation to be the cause of the case.

A diagnosis makes the distinction visible. The clinician observes a set of features and says, in effect:

> This person resembles the class of people represented by diagnostic category D.

The category then indexes accumulated knowledge:

$D \rightarrow$ {common course, common risks, common comorbidities, common treatment responses}.

That can be practically important. The inflation occurs when

resembles category D

quietly becomes

D explains this person.

Generalize it.

The same shift can occur outside diagnosis. A personality score can begin as a summary of responses and end as an explanation of why the person acts. An intelligence score can begin as standardized performance and end as a statement about the person's cognitive essence. A heritability estimate can begin as a property of population variance and end as a statement about what portion of the person was biologically determined.

The practical usefulness of a representation helps obscure the shift because the same term can perform two different jobs. It can identify a recurring pattern. It can also appear to name the mechanism producing that pattern. The classificatory procedure can establish the first. The second requires separate evidence.

Treatment selection shows why the distinction matters even when the practical inference is sound. Suppose a research literature shows that people in class C improve, on average, with treatment T. A clinician encountering a new person who fits C then has reason to consider T:

$C \rightarrow$ evidence that T is worth trying.

That does not establish:

$C \rightarrow$ causal explanation of the person's problem.

A treatment can work across causally heterogeneous cases. If a breathing protocol reduces panic in many people, its efficacy does not show that every recipient's panic arose from the same explanatory history. If exposure reduces avoidance across many cases, successful exposure does not show that one psychological cause generated every avoidance pattern.

Case formulation moves closer to the individual because it asks a different question from diagnosis alone.

State the new problem.

Why does this person have this condition, in this form, at this time?

A formulation may include developmental history, current relationships, beliefs, habits, reinforcement patterns, trauma, coping strategies, physiological vulnerabilities, and environmental stressors. Yet its explanatory parts can still be imported from population-derived constructs such as avoidance, attachment style, core beliefs, rumination, maladaptive coping, trait anxiety, impulsivity, or perfectionism. The case is then individualized by recombining standardized pieces.

That is not necessarily an error. Shared mechanisms do exist.

What licenses the transfer from a shared description to a shared causal structure?

Population knowledge can apply to an individual case only when the relevant mechanism is sufficiently stable across the people grouped together for the purpose at hand.

The implied sequence is:

$$\text{similar observable} \rightarrow \text{same construct}$$

$$\rightarrow \text{sufficiently similar causal mechanism} \rightarrow \text{population evidence applies.}$$

The measurement rule may establish the first transition. The practical inference depends on the last. The step that still requires a causal warrant is the one between them:

$$\text{same construct} \rightarrow \text{same relevant causal mechanism.}$$

Name that missing structure.

Call that transition a claim of **local psychological homogenization**: members need not be psychologically identical generally, but must be causally homogeneous with respect to the mechanism being invoked. It does not say that all people are alike. It says that, for this explanatory or practical purpose, differences among them can be ignored because the category has captured the causally relevant commonality.

Sometimes that may be justified. A tightly characterized visual deficit may arise from a shared lesion. A specific pharmacological mechanism may be sufficiently invariant. A well-specified learning mechanism may transfer reliably. Broad personality constructs, diagnostic classes, and socially saturated behavioral categories leave more room for causally different lives to satisfy the same observational criteria.

Generalize the problem.

A result R_P is established from variation among many people, while the practical claim concerns individual i:

$$R_P \Rightarrow R_i.$$

The problem of supplying such a warrant will be called the problem of **transportability**.

What makes that implication valid in this case?

One reason the warrant cannot be assumed is that a relation among people and a process unfolding within one person are different objects. For example,

$$\mathrm{Cov}_{persons}(X, Y) > 0$$

does not imply

$$X_i(t) \rightarrow Y_i(t+1).$$

Explain.

If people differ in their dynamics, or if those dynamics change over time, a between-person average need not describe the dynamics of any particular individual. The formal conditions under which a structure estimated across persons also describes the structure unfolding within persons are discussed under the problem of **ergodicity**. The plain-language point comes first: averaging across people does not automatically recover the process occurring within one person.

That distinction exposes three separate transfers:

between-person structure → within-person structure,

then

within-person statistical structure → causal structure,

then

present causal structure → future stability despite learning and criticism.

Ergodicity covers mainly the first transition. Causal inference concerns the second. The third transfer, from a present causal model to its future stability in a knowledge-creating person, completes the **three-transfer problem**.

Successful classification establishes none of these automatically.

What happens next?

2. You become the explanatory object

The transport problem changes character once you are no longer treated as the endpoint of the analysis but as the object that still needs explaining.

> Consider a psychologist observing a sustained conversation characterized by recursive criticism, abstraction, concern with classification, insistence on conceptual consistency, and repeated interrogation of the assumptions underlying psychological research.

From those observations, familiar hypotheses are available: high openness, intellectualization, obsessionality, perfectionism, rumination, need for cognition, or unusual autonomy needs.

Now flip the context:

> The conversation is part of a deliberate test of a software system the participant invented, specifically to observe how that system behaves when exposed to an intellectual environment that does not inherit from the same intellectual lineage.

The outward behavior has not changed. The problem situation has.

Many of the classificatory interpretations can become causally unnecessary because the behavior is now intelligible as part of a particular project.

Before that information:

$$O \rightarrow \text{candidate trait or diagnosis.}$$

After it:

$$O \leftarrow \text{software-testing problem.}$$

This is stronger than the claim that context matters.

The added information can remove the need for an entire class of psychological explanations while leaving the observable behavior untouched. The same outward pattern is therefore compatible with radically different causal histories.

The categorical interpretation can be preserved by saying:

> You wrote the software because you possess the very traits I identified.

But now the counterexample has been absorbed into the category.

Abstract the structure.

The invention becomes evidence of the trait, and the trait becomes the explanation of the invention:

novel achievement → trait classification → trait invoked as cause of novel achievement.

What disappears is the content of the achievement. What problem was being solved? Why this architecture? Why this intellectual lineage? Which prior explanations failed? Which criticism produced the new design? Why this solution rather than another?

The trait vocabulary does not answer those questions. It restates the creator using terms abstracted from population regularities. Novelty has been explained by commonality.

Once you are the explanatory object, a different sequence becomes necessary. The inquiry has to recover the particular difficulty as you encountered it, the understanding through which it became a problem, the possibilities then available, the tentative solutions attempted, the criticisms that altered them, and the organization that resulted.

That whole concrete configuration—the circumstances, your understanding of them, the knowledge already available, and the possibilities you can conceive—will be called your **problem situation**.

The same external event can therefore constitute different psychological problems for different people. The event enters a different history of explanations, expectations, skills, habits, and unresolved difficulties. What it means for you, and what responses become possible, depends partly on your history.

Compress it.

An individual-first explanation can be compressed as:

particular problem situation → relevant knowledge history

→ candidate causal explanation → criticism against the case.

This differs from a classification-first sequence:

individual observations → population category → likely mechanism → protocol.

The contrast is not that one sequence uses general knowledge and the other rejects it. The difference is where the causal claim is earned. In the first sequence, general findings become candidates whose relevance has to survive criticism against the case. In the second, the category can too easily stand in for the mechanism.

Individuality now becomes a **positive explanatory problem** rather than the **residue left after regularities have been extracted**. To explain you is to reconstruct how a particular structure came to exist: which problems became salient, which explanations were retained, which criticisms altered them, and which changes reorganized what you could subsequently perceive or do.

Not all of the relevant knowledge need be available as a sentence you can report. Some appears in habits, expectations, skills, sensitivities, or ways of interpreting a situation without being fully articulated. Knowledge that operates in this way will be called **inexplicit knowledge**. Its importance is causal, not mystical: conduct can depend on retained capacities and expectations that are expressed in action without first being stated as propositions.

Neither explicit self-description nor an observed inexplicit regularity is automatically an explanation. Both are evidence from which a causal structure may be conjectured and criticized.

You, on this account, are neither an opaque exception nor a bundle of measurements. You are a **historically developing causal organization** whose present responses

depend partly on the problems encountered, the explanations and habits retained, the criticisms that changed them, and the possibilities you can now conceive.

Generality does not disappear. It changes location.

What generalizes need not be the psychological content of every person. The recurring element may instead be the explanatory method: recover the problem, reconstruct the relevant knowledge, propose causal dependencies, expose them to criticism, and revise them when the case resists.

Shared cognitive and biological mechanisms remain part of the explanation's background. Two people may share language competence, memory architecture, perception, and many physiological capacities while interpreting the same criticism in opposite ways because their explanatory histories differ. The shared mechanisms make many responses possible. They do not by themselves determine the individual meaning of the event.

Nor does individuality imply that you require a private science with no recurring mechanisms. Recurrence has to be established at the level at which the causal claim is made. A population category can suggest where to look. Satisfying its observational rule does not by itself settle what is operating in you.

Right. So population knowledge like classification, scales, averages, traits, diagnoses, risk estimates and generalization moves fast, but knowledge of individuals, the causal reconstruction of their history, problem situation, the mechanisms of conjecture, criticism, and revision moves slow.

Return to the population knowledge and restate within this framework.

3. Returning to population knowledge

With you now treated as the explanatory object, familiar population findings can be reconsidered without discarding what they actually establish.

Suppose a study finds that measured variation in parenting explains relatively little variation in some adult characteristic. That result answers a question about differences within the studied population: how much of the observed variation among people is associated with measured differences in parenting? It does not by itself answer the larger causal question of how parenting contributed to producing the phenomenon at all.

A cause can be nearly constant across a population and therefore explain little variance within that population while remaining indispensable to the phenomenon.

Language makes the distinction clear. Among native Japanese speakers, exposure to Japanese may explain almost none of the variance in who speaks Japanese. Yet Japanese linguistic culture remains central to explaining why they speak Japanese at all.

Compare that structure.

The distinction is:

$$\text{explanation of residual variation} \neq \text{explanation of the phenomenon itself.}$$

An imagined society makes the consequence concrete. Suppose "Agreeablestan" transmits, through language, institutions, norms, parenting, peers, law, ritual, and education, an unusually strong cultural system of agreeable conduct. A child whose genotype would weakly predict lower agreeableness in a Western dataset might nevertheless develop into an entirely ordinary Agreeablestan citizen. Even if subtle biological differences remained, the developed social and psychological organization could be overwhelmingly shaped by common cultural knowledge.

A behavioral-genetic result could still accurately describe residual differences among members of that population. What it would not contain by itself is an explanation of you as a whole. Explaining variation after common developmental structures have done much of the work is not the same as explaining those structures.

This distinction also clarifies the recurrent qualifications attached to heritability claims. Statements such as personality is heritable, intelligence is heritable, or particular traits have genetic correlates are often accompanied by reminders that heritability is not determinism, genes are not destiny, environment still matters, and population statistics do not determine individuals. Those qualifications matter because the underlying quantities concern relations in populations, not a partition of the causal composition of a person.

The slippage appears when a population relation is summarized in language that sounds like a claim about what the trait itself is. "Genetic basis of personality" sounds like a statement about the constitution of personality. "Genetic covariance with variance in a psychometric phenotype" states a narrower relation. The second says less because the evidence establishes less.

No hidden ideological motive is needed for the stronger reading to appear. Once a research program provides

$$\text{trait} \rightarrow \text{heritability} \rightarrow \text{associated loci} \rightarrow \text{polygenic prediction},$$

a parent, clinician, institution, or policymaker can begin to treat the represented trait as though its future expression were already partly contained in the genetic description. The practical form of the result makes that interpretation easy to imagine even when the population evidence does not license it.

Classification has another property that becomes visible only after returning from the person-level inquiry: the representation can enter the causal system it represents. A score, diagnosis, or risk classification can affect access to treatment, educational placement, disability status, insurance decisions, institutional credibility, legal judgments, employment, and self-conception.

A diagnosis can provide relief and useful language while also becoming part of identity. A trait can provide concise description while also becoming a presumed limit. A genetic association can describe population variance while also becoming a story about what a child is destined to be.

People can encounter the theory applied to them, incorporate it into self-understanding, resist it, exploit it, or have their opportunities reorganized by institutions that use it. The classification can therefore become one of the causes shaping what happens next.

That reflexivity is methodological as well as ethical. If applying a representation changes the self-understanding or institutional conditions of the person being represented, a later causal explanation may have to include the effects of the representation itself.

Nothing in this argument removes the practical value of population knowledge. It narrows the kind of claim that has been established. Population evidence can reveal risks, common responses, useful interventions, recurring patterns, and possibilities that deserve attention. By itself, it does not complete the causal explanation of the person to whom the evidence is applied.

Now point that population knowledge at the individual.

4. Individual explanation under criticism

Turning to you creates a symmetrical danger. If your case is treated as so particular that no general standard can touch it, explanation collapses into storytelling. Any narrative could be protected from criticism by declaring you unique.

An explanation of you therefore has to expose its causal commitments. It needs to identify the problem being solved and the relevant knowledge, including both what you can state and what appears through habits or expectations. It also has to specify the proposed dependencies, the observations those dependencies account for, the counterfactuals they imply, the evidence that would defeat them, and a rival explanation capable of replacing them.

Neither intimacy nor classification supplies causal authority.

There is no such thing as causal authority. Restate the mechanism.

A clinician's statement that they understand you does not establish an explanation. Your own self-description may be corrected. A population category does not become a mechanism merely because you satisfy the category. All three remain open to criticism.

Several established problems in psychological inference fit inside this reciprocal demand. Relations estimated across many people need not describe the process unfolding within you; this is the familiar contrast between nomothetic and idiographic description and the formal problem of non-ergodicity. But intensive individual measurement does

not finish the job either. A high-resolution time series remains a statistical description until its causal interpretation is independently explained.

Stop! We're trying to explain *you*, not psychology! Focus on the constraints.

Variation across situations adds another constraint. Stable tendencies can appear as conditional patterns rather than invariant outputs. But for you, the relevant situation is not always exhausted by the external stimulus configuration. The **same event can become a different problem when different knowledge makes it intelligible in a different way**.

Historical and reflexive approaches add a further complication. Psychological regularities can change when culture changes, and psychological knowledge can itself alter behavior. At the level of your own case, the causal point is that you can acquire, criticize, and create knowledge. **New knowledge can reorganize the system that generated the earlier behavior**.

Statistical sophistication does not remove that explanatory burden. Statistics can perform excellent error elimination while leaving open what causal structure a construct name refers to. The problem is not quantification. It is **using quantification to answer a causal question before the explanatory object has been established**.

These criticisms also have limits. Descriptive regularity does not disappear merely because a trait is not a deep causal primitive. Cultural explanation does not imply a biological blank slate. Ordinary genetic variation may alter affect, attention, arousal, perception, or other conditions under which problems arise. The narrower claim is:

> Biological difference cannot be promoted into a limit on explanatory possibility without a mechanism establishing that limit.

Nor does an individual-first approach license unaided clinical intuition. Statistical prediction can outperform unaided professional judgment. Population evidence can still generate possibilities, identify risks, and prioritize interventions. The distinction concerns inferential role, not whether population knowledge is allowed into the case.

Historical or social construction does not imply unreality either. A psychological category can be historically constructed and still track something causally real. Every scientific category is invented in the sense that the classification is made by people. The relevant question is whether the proposed causal structure can be altered freely without losing what it explains. A structure that cannot be varied in that way will be called here a **hard-to-vary explanation**.

Good enough. But we need some structure to carry it. Compress these ideas algorithmically.

5. Four tests

The argument can now be compressed into four tests. Each asks for a warrant that either population classification or person-specific explanation can fail to provide.

First: has a relation estimated across people actually been shown to describe a process within you? This is the **Ergodicity Test**:

> When population evidence is applied to you, what establishes that the relevant between-person relationship corresponds to a process within you?

Without that demonstration, the population structure remains population evidence rather than an individual mechanism.

Second: has a name for a recurring statistical pattern been mistaken for the causal structure that generates it? This is the **Reification Test**:

> When a psychological construct names a recurring covariance pattern, what independently established causal structure does the construct refer to?

A stable factor solution is not itself the answer. The construct may summarize the observations successfully while leaving open what produces them.

Third: could an apparently stable limitation be reorganized if you acquired an effective criticism of the knowledge producing the behavior? This is the **Epistemic Plasticity Test**:

> What would have to be true for explanatory revision to reorganize this pattern, and what mechanism would prevent that reorganization if the constraint is claimed to be fixed?

A claim of fixed constraint therefore carries a causal burden. Stability observed so far does not establish stability under knowledge not yet acquired.

Fourth: what could show the account of you to be false? This is the **Idiographic Discipline Test**:

> What observation or counterfactual would defeat the explanation?

If no possible finding could count against the account, individuality has become a shield against criticism rather than an explanatory object.

The four tests impose reciprocal discipline. Population models have to earn individual applicability. Individual explanations have to earn causal status.

The contrast is therefore not statistics versus stories, quantitative versus qualitative, general versus unique, or objective versus subjective. The relevant distinction is between

representations whose evidential reach is known and explanations whose causal claims have survived criticism at the level where they are being applied.

I find it ironic that we're categorizing and systemizing the thing that we first criticized as categorical and systemic. But let's wrap this up with a tentative solution to our problem.

6. What can population knowledge reveal about you?

The inquiry can now return to its starting point. Conventional psychology produces knowledge about population differentiation, behavioral measurement, classification, prediction, intervention efficacy, and institutional decision support. Those are substantial achievements and can produce useful, sometimes life-improving technologies.

The explanatory error begins only when success at those tasks is treated as though it automatically supplied a causal account of you.

Once you enter as the explanatory object, two questions that had been allowed to collapse into one have to be kept distinct:

> Which known structure do you instantiate?

and

> What structure is actually operating here, how did it come to operate, and what would show that we are wrong about it?

The first question can narrow possibilities, surface risks, and connect the case to accumulated knowledge. The second asks whether any of those possibilities actually explain you. Classification can begin that investigation. Naming the resemblance cannot complete it.

The deeper contrast is therefore not between generality and uniqueness. It is between **classification-first explanation** and **individual-first explanation**.

Classification-first explanation is fast. It begins with a representation whose meaning was stabilized across a population and asks how you fit it. Individual-first explanation is slow. It begins with your actual problem situation and asks which general findings, if any, survive criticism when transported into your case.

That change of direction also changes what population knowledge can mean personally. A population result may show that people resembling you in some measured respect have often followed a certain course, responded to a certain intervention, or displayed a certain risk. That is information about possibilities that may deserve attention. It is not yet an explanation of why you are as you are.

A causal explanation requires the organization particular to the case to be reconstructed: the relevant problems, the knowledge through which they are understood, the

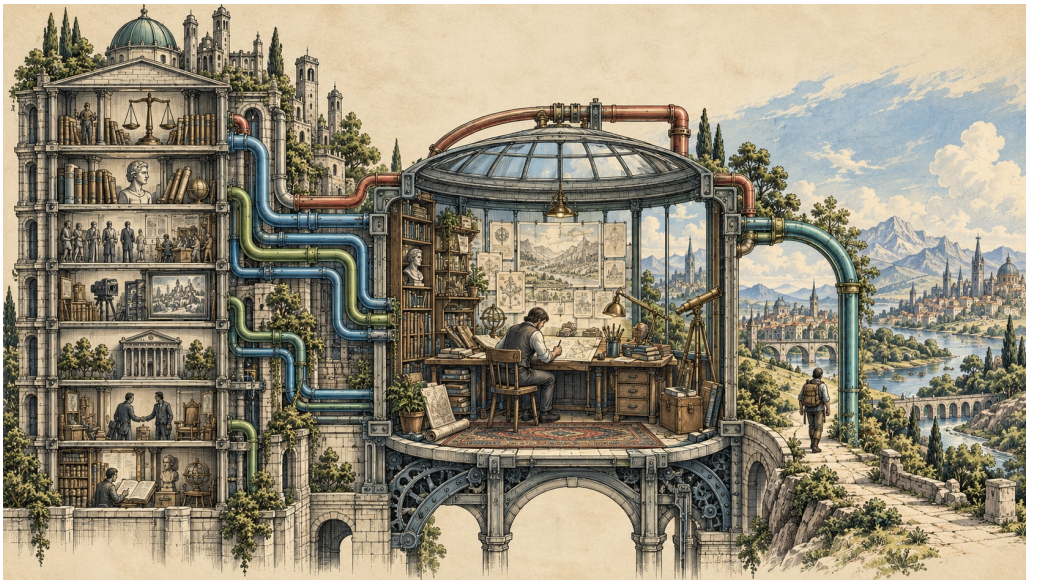
explicit and inexplicit expectations that shape action, the criticisms already encountered, and the changes that remain possible.

That reconstruction illuminates the next problem. If you can acquire criticism, create new explanations, and reorganize the knowledge through which a situation is understood, then the object being explained is not merely particular. You can change the very structure from which your previous regularities were generated.

The remaining question is therefore harder than how to model an individual more accurately:

> How can a causal explanation of you remain testable while taking seriously that you can acquire new knowledge that changes the causal organization being explained?

III. The Anxious Authorities



At the social level, the problem of *you* changes because society is not merely a distant influence. It supplies much of the material with which you think: language, rules, roles, expectations, explanations, institutions, laws, stories, technologies, and inherited ways of framing problems. Call it **social objective knowledge**, or social knowledge.

Social knowledge can therefore act on you far more directly than a statistical relation between genetic variation and behavior. That makes the social case both stronger and more dangerous.

The error is not that society has little causal effect. The error appears when knowledge about a social regularity is treated as knowledge of the kind of person who participates in it.

Abstract that idea.

The conventional movement is familiar:

$$\text{group regularity} \rightarrow \text{group property} \rightarrow \text{individual explanation}.$$

A bureaucracy exists. A legal system exists. A language exists. A political doctrine exists. The social object can be real, and the regularity measured across it can also be real. But the reality of the object does not make the category naming it an explanation of the individual who encounters it.

Locate the pattern in politics.

A political survey makes the distinction unusually clear. Suppose two people both endorse proposition Q:

$$P_1 \rightarrow Q$$

and

$$P_2 \rightarrow Q.$$

The survey may have discovered something straightforward: both people presently endorse the same proposition. What it has not discovered is why either person endorses it. One may have reached it through moral reasoning, another through personal experience, economic argument, trust in an authority, or failure to consider an alternative.

Compare it.

The important relation is:

$$\text{same expressed knowledge} \neq \text{same explanatory history}.$$

The group description can represent something real while leaving the person unexplained.

The social case goes further because many social objects are themselves shareable structures that survive any one person's participation in them. A language can be learned by different speakers. A constitution can be read, criticized, amended, obeyed, or ignored by different citizens. A professional code can constrain people who disagree about its justification. A moral tradition can be inherited before you can state the principles it contains.

Call these publicly available structures **objective knowledge**: knowledge embodied in forms that can be evaluated across solitary minds. You do not begin by inventing the language, law, institution, or tradition. You encounter it already there.

That encounter changes what becomes thinkable and actionable. A rule can close an option. A norm can make one action intelligible and another shameful. A hierarchy can attach sanctions to disagreement. A doctrine can supply an explanation before you have encountered a criticism of it.

A gene cannot tell a child that adults know better. A culture can. A gene cannot tell an employee that managers are entitled to decide. An institution can. A gene cannot tell a citizen that obedience to authority is virtuous. A political tradition can.

Society therefore does not simply surround you. It helps constitute the field of problems, expectations, explanations, and constraints within which you act.

Call that surrounding structure your **problem situation**: the problems you can see, the knowledge you have inherited, the options you believe are available, the sanctions you anticipate, the explanations you can presently use, and the criticisms you can presently formulate.

If social knowledge enters your problem situation, what explains your action?

Take a bureaucrat. At work, the person may be formal, procedural, risk-averse, deferential to hierarchy, and emotionally restrained. At home, the same person may be playful, spontaneous, irreverent, affectionate, and improvisational.

Labels such as role, organizational culture, situational norm, or identity activation can mark real regularities. But they do not yet tell us what this person understands themselves to be doing.

The person may understand the workplace as a system in which procedure defines admissible action, disagreement carries career costs, success means avoiding unauthorized risk, and superiors possess legitimate decision rights. At home, those problems may disappear. Other obligations, possibilities, and expectations become active.

Abstract that structure.

The causal structure is closer to:

$$\text{socially inherited knowledge} + \text{individual problem situation}$$

$$\rightarrow \text{interpretation} \rightarrow \text{action}.$$

The same mind can behave differently because different knowledge and different problems become active in different circumstances. No contradiction in personality is required.

This does not make the institutional regularity unreal. If an organization gives many people similar procedures, incentives, sanctions, vocabularies, and decision rights, similar behavior is unsurprising. The regularity may explain something important about the institution. The point is narrower: the causal route into action still passes through a person who interprets what those conditions mean.

Nor is interpretation unconstrained. Institutions can alter the actual possibility space around you. Law changes which actions are permitted. Censorship changes which criticism can circulate. Employment hierarchy changes who possesses formal decision rights.

Social conditions can also remove conditions needed for the normal exercise of explanatory agency. Solitary confinement, for example, can remove ordinary social interaction and contribute to psychological deterioration. That does not make the mind a lesser kind of mind. It shows that explanatory agency has enabling conditions: a universal capacity can be impaired when conditions required for its exercise are removed.

Compare it.

The distinction is therefore not between social causation and individual freedom. It is between two different explanatory claims:

$$\text{restricted action space} \neq \text{restricted explanatory capacity}.$$

You may act cautiously because criticism is punished. An employee may defer because formal authority is backed by sanctions. A bureaucrat may follow procedure because

procedure determines what counts as an admissible move. Those conditions can explain recurring behavior without becoming an intrinsic property of the people who behave that way.

This is where social essentialism enters.

Stop. There's no such thing as an essence. Reverse the definition.

Behavior produced under conditions is redescribed as a property of the people who inhabit them: bureaucrats are rigid, members of hierarchical societies are deferential, children are suggestible.

Compress it.

The observation may summarize a distribution. The explanatory inflation occurs when:

behavior under conditions

becomes

property of person.

A social category can tell us which roles exist, what sanctions are likely, which propositions are widely shared, or which actions are formally possible. It does not by itself tell us which of those things this person accepts, fears, misunderstands, criticizes, combines, or rejects. That requires your explanatory history.

Much of that history need not be explicit. A child can learn who is allowed to interrupt whom before possessing a theory of hierarchy. An employee can learn which disagreements are career-limiting without formulating a doctrine of managerial authority. A student can learn which questions count as challenges before being able to describe the rule being enforced.

Because this is knowledge rather than essence, it can conflict with other knowledge. You can retain one rule in one domain and reject it in another. An inherited norm can remain invisible until criticism exposes it. An explanation can be revised without you becoming a different kind of mind.

Knowledge learned this way can remain inexplicit: operative in action before you can state the rule being followed.

You are therefore not merely the endpoint of social causation. You are where inherited social knowledge is interpreted, combined, criticized, and sometimes changed.

Authority makes the structure especially visible.

A persistent social pattern can be understood as a network of propositions, norms, habits, and institutional arrangements concerning who is entitled to decide for whom. It need not be a collective mind or a hidden social trait. Its content may be learned through recurring relations: parents over children, teachers over students, managers over employees, experts over non-experts, rulers over citizens.

These relations do not merely recur independently. They can persist as inherited knowledge, reproduced through explicit propositions, inexplicit habits, recurring practices, and institutions. Call such a transmissible unit of social knowledge a **meme**. Where several such memes support and reproduce one another, they form a **memeplex**. The recurring knowledge concerning who is entitled to decide for whom can therefore be treated as an **authority memeplex**.

Abstract it again.

The relation can become familiar before anyone states it as a doctrine:

$$\text{higher-status actor} \rightarrow \text{greater decision authority}.$$

Modern liberal societies also inherit a newer memeplex organized around autonomy, equality, freedom of speech, democratic participation, freedom of conscience, informed consent, and individual rights. This newer knowledge can coexist with the older authority memeplex inside the same society and inside the same person. The newer vocabulary can therefore sit on top of older authority relations without fully replacing them. The older structure may reappear when applying the newer principle would dissolve an established authority relation.

One relation says:

$$\text{person} \rightarrow \text{presumptive autonomy}.$$

Another says:

$$\text{authority differential} \rightarrow \text{legitimate control}.$$

You may defend freedom of judgment in politics while treating obedience as obvious at work. An employee may resent managerial control and reproduce the same authority relation with a child. A professional may affirm autonomy while regarding expertise as sufficient grounds to override another person's interpretation.

A hidden trait called authoritarianism is not required to make those conflicts intelligible. Partially incompatible bodies of inherited knowledge may become active in different problem situations.

The personal question is then, which knowledge is active here, what problem does it appear to solve, and what criticism could expose the conflict?

Children put unusual pressure on this question because differences in knowledge and practical competence are obvious.

The familiar argument is:

> children know less, therefore adults should decide for them.

The first proposition is often true. The second introduces a stronger relation:

less knowledge → less epistemic sovereignty.

Once stated in that form, the principle is not logically confined to childhood. It can be extended wherever one person or institution judges another to know less: patients, students, employees, voters, or populations regarded as vulnerable to misinformation.

When differences in knowledge are used to justify differences in your authority over your own judgment, the result is a doctrine of **graded epistemic personhood**.

The child case is relatively easy because unequal knowledge is visible. But unequal knowledge and unequal kinds of mind are not the same claim.

A pilot may know more than a passenger about landing an aircraft. A doctor may know more than a patient about a disease. An adult may know more than a child about traffic. These differences concern knowledge possessed in a domain. They do not establish a different capacity for creating knowledge.

Abstract it.

The distinction is:

knowledge possessed

versus

capacity for knowledge creation.

A child begins with less knowledge, fewer experiences, fewer available explanations, and less practical control over circumstances. Those are real features of the child's problem situation. Yet the child still has to interpret what happens. The child forms expectations, notices anomalies, invents explanations, imitates, asks questions, misunderstands, tests, criticizes, and revises.

Authority can therefore change more than what the child is allowed to do. It can change which problems the child is permitted to pursue, which explanations are available, and which criticisms survive long enough to develop.

Media makes that change easy to see.

One common model treats the child as a passive recipient. The medium acts; the child absorbs. Its causal structure is:

content → mind.

We will call this **the bucket theory of mind**. Once the child is pictured as a container, controlling inputs looks like a direct way of controlling what enters the mind.

But the child cannot receive a medium without interpreting it. A child looking at a screen must conjecture what is happening, what matters, what another person intends, what will happen next, and which rules govern the situation.

Abstract it again.

The structure is closer to:

media environment → problems encountered → child's conjectures and criticism.

Removing a screen therefore does not merely remove content. It may terminate an active problem environment selected by the child. A parent may imagine that "go outside and play" moves the child into a richer world. The child may experience the opposite: a move from a dense unresolved problem space into one containing nothing presently salient.

The physical medium does not settle that difference. The explanation depends on what problem this child is pursuing through it.

That same distinction helps with recurring concern about new media. Reading, radio, television, the internet, social media, and generative AI differ technically and can generate different harms. Any specific criticism may turn out to be correct. But as one medium becomes ordinary infrastructure, concern can migrate to the next. That recurrence does not show that the specific criticisms are false. It suggests that the changing technical criticism may not exhaust the explanation. Each medium also changes the routes by which you encounter knowledge and who or what controls your next epistemic encounter.

Social media changes association. A rare interest once constrained by geography can become a searchable community:

rare interest + geographic dispersion → searchable community.

That same structure can support ideological reinforcement, specialist collaboration, support networks, or scientific exchange. "Polarization" can describe one outcome without exhausting the architectural change.

Let's introduce AI as a new medium and contrast it with older forms of media.

Generative AI changes something else. Social media helps you find others already organized around an interest. A search engine retrieves from an existing informational landscape. A generative system can recombine inherited knowledge around a problem that may not yet correspond to an existing document, course, community, or established query. It can therefore create a responsive interaction around your present problem:

individual problem → generative system
→ candidate explanation → criticism → new problem.

Call it an **application of epistemic processing**.

For a child pursuing an unusual scientific, technical, philosophical, or artistic problem, the former bottleneck may have been the accidental availability of an adult or institution able to respond. A generative system can alter that bottleneck by making a candidate explanatory response immediately available.

That does not establish that the response is true, reliable, or worth retaining. It changes the problem situation. It changes what can be asked next and how quickly a candidate explanation can be produced, criticized, and replaced.

This is why opposite complaints about media need not identify one mechanism. Social media can be criticized for excessive differentiation; AI for excessive commonality; unusually capable AI for threatening human distinctiveness. Those claims concern different possible effects. What the media can share is a transfer of some **control over the next epistemic encounter** away from incumbent curators, away from the incumbent authorities.

And the authorities become anxious.

That transfer is not, by itself, a judgment about whether the change is good or bad. It identifies a change in the distribution of epistemic control.

Now return to the original social-scientific problem. This personal analysis changes its significance.

A theory about society can become part of society. A claim that children are vulnerable to media can alter parental behavior, school rules, platform policy, regulation, and public norms. Claims that workers require supervision, voters are irrational, social media polarizes, or AI causes dependence can likewise be built into institutions.

Once institutionalized, such claims cease to be mere descriptions. They become causal inputs into the lives of the people described.

Abstract it.

The sequence can be:

$$\text{population observation} \rightarrow \text{social category} \rightarrow$$
$$\text{policy interpretation} \rightarrow \text{institutional rule} \rightarrow \text{individual problem environment.}$$

Suppose an association between media use and distress in a population of children is redescribed as "children are vulnerable to media," and that description then supports rules controlling children's access. Whether the policy works is a separate empirical question. The epistemic transition has already occurred: a class-level observation has been translated into an authority relation over particular people.

The social sciences therefore contain a reflexive causal route:

$$\text{theory of society} \rightarrow \text{policy} \rightarrow \text{social structure} \rightarrow \text{individual problem environment.}$$

A weak explanatory category can become a strong institutional constraint.

That return also clarifies what the original social knowledge can genuinely tell us. Social categories can identify distributions, shared propositions, institutional rules, recurring behavior, authority structures, legal restrictions, communication networks, and changes in information environments. Some social structures provide direct causal

explanations of constraints: a law prohibits an action, a hierarchy allocates a decision right, censorship removes a channel of criticism.

What the category does not contain merely by naming the group or regularity is the explanatory history of the individual.

The conventional movement tends to be:

$$\text{group} \rightarrow \text{regularity}$$
$$\rightarrow \text{social category} \rightarrow \text{individual expectation} \rightarrow \text{policy}.$$

The epistemic movement keeps the social knowledge but relocates you inside the causal chain:

$$\text{socially inherited knowledge} + \text{individual problem situation}$$
$$\rightarrow \text{interpretation} \rightarrow \text{criticism or retention} \rightarrow \text{action}.$$

Society changes the concepts available to you, the problems that become salient, the explanations encountered before criticism begins, the sanctions attached to dissent, the criticism that can circulate, and the actions that are possible. You may accept that inherited knowledge, misunderstand it, combine incompatible parts of it, reproduce it elsewhere, criticize it, or replace it with a new explanation.

That is how the same social knowledge can produce regularities without producing identical minds. It is also how you can participate in a regularity without being explained by the category that records it.

The deeper distinction is not simply between a group statistic and an individual case. It is between social knowledge and the explanatory history of the person who encounters that knowledge. The first can describe real institutions, propositions, constraints, and distributions outside you. The second explains how you encounter that world and what happens next.

Confusing them turns social knowledge into a substitute for you.

The alternative is not to imagine yourself outside society. Societal knowledge is close to the content of individual minds precisely because it is inherited through language, institutions, norms, laws, roles, media, technologies, and authority relations. It can constrain action, alter which criticism is available, and determine which problems become visible first.

Its causal power is why classificatory overreach matters. But none of those transitions requires a lesser kind of person beneath the constraint.

I feel an abstraction coming on.

The negative distinction remains:

$$\text{social constraint} \nRightarrow \text{lesser explanatory person}.$$

The positive relation is more important:

socially inherited knowledge

↓

person-specific problems

↓

possible criticism and new knowledge.

You can inherit norms of autonomy and hierarchy at the same time. A child can know less without being a lesser kind of knower. A medium can change a child's epistemic landscape without mechanically writing content into the mind. An institution can restrict action without defining the explanatory capacity of the person subject to it.

The social world enters you as a field of inherited explanations, constraints, and problems. It does not enter as a completed psychological essence.

There's no such thing as an essence. What's the next problem?

If social knowledge acts through individual problem situations, where is the society that appears to know, decide, and command?

IV. The Wisdom Between The Crowd



You remain the locus of explanation. Here, that commitment is applied to society itself.

Wait. That sounds like a contradiction. The individual is the locus of society?

The conventional language of social life is saturated with collective subjects: society believes; the government knows; the market decides; science says; the church teaches; the company wants; the public demands. These expressions are useful shorthand, but the shorthand can become an ontology.

Speak clearly.

A stable social pattern begins to sound like a subject that sees, judges, intends, and knows. There is no social mind in the explanation developed here. There are individual minds. Each person is a distinct locus of experience, conjecture, interpretation, criticism, and action. Each knows only part of what other people know and only part of the social world in which they act. Communication can connect these minds without merging them. Institutions can coordinate them without creating a superordinate knower.

Explain.

What persists between people is not a shared subjectivity but knowledge made objective. Call it **objective knowledge**. Explanations, rules, classifications, procedures,

records, software, texts, and standards can outlast the persons who created them and later be reconstructed by others. Such structures can have enormous causal effects without possessing experience of their own.

A law can constrain you without knowing you. A holy book can reorganize a life without possessing a mind. A bureaucracy can preserve procedures without understanding them. An artificial intelligence system can transform and recombine objective knowledge without thereby establishing that it is a knowing subject.

The problem is therefore narrower than a complete theory of cultural transmission, collective coordination, or machine cognition. Two explanatory frames are in conflict. The conventional frame permits societies, institutions, technologies, and fields of knowledge to appear as superordinate epistemic actors.

Abstract it.

The epistemic cycle requires every such appearance to decompose into the same universal process of knowledge creation:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2$$

Problem, tentative solution, error elimination, and new problem, instantiated in the minds of individual explainers.

If society requires a special collective process of knowing, this epistemology is not universal and therefore false. Applying the explanation recursively is therefore not an optional extension of the theory. It is a consistency test.

Where do we begin?

1. The solitary standpoint

Knowledge can be objective. Experience cannot.

A scientific paper may exist independently of any present reader. A law may remain in force after everyone who drafted it is dead. A computer program may encode knowledge that no current user could recreate. An institution may preserve procedures across generations.

But every encounter with any of these structures still occurs from one subjective standpoint.

You do not experience society as a whole. You experience particular conversations, rules, media, institutions, sanctions, documents, and other people. From those fragments, you construct an understanding of a social world that is never directly available in its entirety.

That understanding is partial, but partiality is only the first difficulty. What reaches you must also be interpreted. You form conjectures about what others believe, what

institutions intend, which norms apply, who has authority, which actions are permitted, what consequences are likely, and what "society" expects.

Those conjectures may be accurate. They may also be wrong.

Nothing unusual has to happen for the error to arise. No class, profession, institution, or historical period has to be especially confused. The possibility follows from the structure of individual experience itself: no person occupies the standpoint of society.

That creates the first explanatory problem. If nobody encounters the social whole directly, how does social life acquire the appearance of unity at all?

Communication seems to offer an answer. One person has an idea, expresses it, and another receives it. Yet that description already assumes the thing that needs explaining: that the same knowledge has somehow crossed from one mind into another.

A speaker has some internal explanatory structure and externalizes part of it through language, gesture, artifact, or action. Another person encounters that expression. The second person does not receive the speaker's mental state. They reconstruct meaning using their own prior knowledge, language, expectations, current problems, and criticism.

Even successful communication therefore produces overlap rather than identity. Two people can use the same words while possessing different explanatory structures. They can also converge on the same proposition through very different histories.

The important result is not that communication fails. It is that coordination does not require mental identity. Enough overlap can sustain a legal system, laboratory, company, church, school, or family across time. Stable social patterns can emerge from many individual minds reconstructing sufficiently compatible models of one another and of the objective structures through which they act.

The appearance of a social whole begins to dissolve. The next question is whether institutions put it back together at a higher level.

2. Institutions without an institutional mind

Institutions can look as though they escape individual ignorance. A government may contain intelligence agencies, statistical offices, ministries, databases, advisers, courts, and bureaucracies. A corporation may contain finance, legal, engineering, operations, sales, and executive divisions. A university may contain hundreds of disciplines.

The larger the structure, the more knowledge it appears to contain. But the knowledge is distributed. No single mind has access to all of it.

The institution coordinates partial perspectives through reports, forms, meetings, databases, procedures, chains of command, software, policies, and other objective structures. These let people who do not share the same knowledge, and may not even know one another, participate in a stable pattern of action.

That produces a functional whole without producing an epistemic whole.

A government may "know" a statistic in the sense that the statistic exists somewhere in its records. That does not imply that any person making a relevant decision has encountered it, understood it, or interpreted it correctly. A corporation may "decide" to pursue a strategy, but the decision resolves into individual judgments, communications, delegated authority, procedures, and recorded actions.

The institution is causally real. What has not appeared is a new subject of experience.

This distinction matters because objective knowledge can act without awareness. A contract can bind an employee. A diagnostic manual can alter how a clinician interprets a patient. Software can block access to a service. An algorithm can route information. An LLM can produce a candidate explanation. None of these effects requires the artifact itself to possess subjective experience.

The distinction avoids two opposite mistakes. If only persons experience, it does not follow that objective structures are causally unimportant. If objective structures have powerful effects, it does not follow that they are minds.

You are still where the structure becomes lived: encountered, interpreted, criticized, obeyed, resisted, or reconstructed.

That return to *you* makes coercion less abstract.

A rule can alter another person's action space through force, threat, punishment, exclusion, loss of resources, legal restriction, surveillance, censorship, or denial of access. An institution can scale the effect, but the causal chain still runs through persons creating, maintaining, or enforcing objective structures that change what other persons can do. A law does not arrest anyone by itself. A school rule does not remove a child from a classroom by itself. A corporate policy does not dismiss an employee by itself.

Constraint can also survive after the enforcing person disappears from the immediate scene. Someone may act within propositions such as: I must obey; people like me should not do this; the expert knows better; this question is inappropriate; this authority is entitled to decide; there is no alternative.

Calling this internalization can make it sound as though "society" has entered you. The more exact sequence is simpler. You encounter socially inherited objective knowledge, reconstruct it, and later use that reconstruction in your own reasoning. The social structure becomes causally effective through your own explanatory system.

External and internal constraint therefore meet at the same point. One changes the available action space. The other can change which alternatives you can see as available at all.

What does this imply about freedom?

3. Freedom, authority, and error correction

Once constraint is described in those terms, freedom acquires a specifically epistemic role.

Error correction requires alternatives. You need some possibility of formulating a problem, considering rival explanations, encountering criticism, testing conjectures through action, and revising conclusions. Coercion can interfere with any of these steps. A censor can restrict rival explanations. A school can forbid a question. A manager can punish dissent. A government can constrain association.

The effect is not exhausted by suffering or domination. Restricting problems, criticisms, associations, or tests can prevent errors from becoming visible and can stabilize false knowledge.

In this inquiry, freedom therefore names more than the absence of restraint. It names a condition in which no other person has unilateral control over the problems you may formulate, the criticisms you may encounter, or the actions through which you may test a conjecture.

That formulation immediately encounters an apparent objection. People do not know equally much about every problem. A patient may rely on a surgeon. A passenger may rely on a pilot. A student may rely on a specialist. Local asymmetries of knowledge are ordinary and often indispensable.

Generalize it.

The difficulty appears when a local asymmetry is transformed into a general rule:

$$\text{less knowledge} \rightarrow \text{less decision authority}$$

The person granted authority remains another partial and fallible mind. A manager may know more about one organizational problem and less about another. An expert may know more about a technical domain and little about the other person's own goals. A government department may possess aggregated data while misunderstanding the local problem of an individual citizen.

Authority can redistribute decision power. It cannot create an omniscient node.

The solitary standpoint therefore survives every institutional level. The relevant question is no longer whether one person sometimes knows more than another. It is what happens when partial knowledge is converted into durable control over another person's problem situation.

Can we make this more specific?

Psychology provides a compact case because the whole social structure can return to one person through a single professional encounter.

4. Who knows you?

Psychology is a field, but more tangibly a vast network of researchers, clinicians, reviewers, editors, students, research participants, journals, databases, statistical conventions, professional bodies, diagnostic systems, and institutional rules. Each participant knows only part of the field. What appears as disciplinary knowledge is distributed across individual minds and objective knowledge.

And knowledge moves.

A theory may originate in one person, be revised by another, encoded in a questionnaire, standardized in a manual, taught in a textbook, implemented in software, and applied by a clinician. At each transition, somebody reconstructs what was inherited or transferred.

This explains how a discipline can display more continuity than any individual mind within it. Knowledge persists because it has been externalized. Without that persistence, each generation would have to rediscover everything.

But persistence is neutral about truth. The same machinery can preserve an error. A questionnaire may survive because it is standardized. A construct may persist because journals recognize it. A diagnostic category may remain because reimbursement systems require it. A statistical method may become conventional because software implements it.

Persistence therefore does not imply continued explanatory adequacy. A field can preserve an inherited problem formulation after the reasons for adopting it have become obscure.

Now return to *you*.

A patient does not encounter "psychology." The patient encounters one clinician, one questionnaire, one report, one diagnosis, one treatment protocol. The clinician carries compressed objective knowledge inherited from a much larger disciplinary network, but the macro structure becomes locally effective in a relation between particular persons.

The diagnosis can then return to the patient as new objective knowledge. The patient may reconstruct it and begin to interpret themselves through it. A theory used to describe a person can therefore become part of the phenomenon it later measures.

That is the reflexive power of social knowledge.

The claim is not that a field should stop classifying, measuring, or preserving knowledge. The narrower point is explanatory. A disciplinary representation can contain information about a person without containing the whole explanatory history of the person who encounters and reconstructs it.

The same distributed structure also explains how claims can expand while moving through a network.

A researcher reports that a measure correlates with outcome X. Another reports a genetic association. A clinician learns that a construct helps classify patients. Later, a journalist says that characteristic X is partly genetic, and a parent concludes that a child is naturally like this.

No single person has to traverse the whole inferential chain. Successive local reconstructions can transform the claim until the public or professional conclusion substantially exceeds the original evidence, even when each participant regards their own step as reasonable.

Isn't that why we have meta-analyses and replicable studies?

Meta-analysis cannot automatically correct this kind of inheritance. Different studies may use independent samples while sharing the same psychometric ontology, trait definitions, measurement assumptions, and statistical framework. Meta-analysis can reduce random sampling error. It cannot, merely by aggregating samples, criticize the common frame inherited by all of them.

Replication across samples is therefore not the same thing as replication across problem formulations.

A field can repeatedly corroborate a result inside one inherited frame while leaving that frame itself largely untouched. When you later encounter the result, you usually do not encounter this history of construction and criticism. You encounter a compressed representation whose apparent authority may exceed the criticism that produced it.

The pattern becomes still clearer when the representation leaves the discipline and becomes policy.

Abstract the structure.

Individual observations can be aggregated into a social category. The category can become policy knowledge, and policy can be implemented through institutions. The institutional action can then return as a constraint on an individual:

$$\text{individual observations} \rightarrow \text{aggregate representation}$$
$$\rightarrow \text{policy knowledge} \rightarrow \text{institutional action} \rightarrow \text{individual constraint.}$$

A category may begin as description and end as a rule governing a person. A population is judged vulnerable and access is restricted. A group is judged irrational and an authority structure is introduced. A demographic category becomes a risk class.

No social mind is required. Researchers, administrators, policymakers, and professionals act through objective structures, each seeing only part of the chain. Yet the chain can become stable enough to act back upon the individual from whom the original category was abstracted.

What does the aggregate contain when it returns?

It may contain regularities, risks, correlations, administrative categories, or predictive information. It does not thereby contain the person's entire problem situation,

knowledge history, interpretations, criticisms, goals, or capacity to create explanations not represented in the category. The aggregate can describe something about a person without becoming the person's explanatory standpoint.

Then the medium really is the message after all?

5. Media and control of the next encounter

The same decomposition changes what "media" means. A medium changes the conditions under which you can encounter, reconstruct, criticize, and continue knowledge.

Explain.

Media are technologies through which objective knowledge is stored, transmitted, selected, or generated. Their history can be read as an expansion of the epistemic functions available to individual minds. Print makes knowledge durable and reproducible. Broadcast lets one communication reach many recipients simultaneously. The internet makes a large body of knowledge individually navigable. Social media reduces geographic and temporal barriers between persons. Generative AI makes objective knowledge responsive to your current problem.

Compress the structure.

The sequence is not a progression toward a social mind. It is a change in how individual minds can encounter objective knowledge.

This becomes important when media effects are compressed into a direct causal shorthand:

$$\textit{medium} \rightarrow \textit{child}.$$

Television influences the child. Social media polarizes the child. AI teaches or manipulates the child.

Abstract it.

The shorthand removes the epistemic middle. A more exact sequence is:

$$\text{knowledge created by persons} \rightarrow \text{technological instantiation}$$
$$\rightarrow \text{routing / selection / generation} \rightarrow \text{child's individual reconstruction.}$$

The medium changes what can reach the child, how quickly it can arrive, how it can be selected, and what forms of response are possible. It does not replace the child's conjectural actions.

The same physical device can therefore mediate activities with very different epistemic structures: passive viewing, conversation, coding, research, argument, art, technical collaboration, tutoring, or problem solving. "Screen time" compresses all of them into one physical exposure variable.

Does this explain our collective anxieties?

Historical anxieties around new media can now be stated more precisely. The specific concern changes: print can corrupt; radio distracts; television makes passive; the internet exposes; social media fragments; AI homogenizes, replaces, deceives, or overpowers. The recurrence does not show that every concern is false. Each technology has real causal properties.

What persists across the transitions is a change in who can control the next epistemic encounter.

Print weakens local oral gatekeeping. Broadcast weakens geographic information boundaries. Search weakens institutional scarcity. Social media weakens the scarcity of rare-interest communities. AI weakens the scarcity of responsive explanation.

The authority problem survives the technological change because the routing of knowledge changes the problems and criticisms you can encounter.

Social media makes one side of this especially visible. The same network relation that can produce polarization under some conditions can also make highly specialized association possible. Rare interests that were previously geographically isolated can become viable communities. What appears as fragmentation at an aggregate level may, from your standpoint, be increased ability to find other people organized around unusually specific problems.

That does not establish that every resulting community is epistemically healthy. It establishes something narrower: an aggregate category such as "polarization" can hide changes in the topology of association experienced by particular persons.

So the medium is not a subject acting directly on a population. It changes the structure of possible encounters. The consequence still occurs when particular people reconstruct what reaches them and carry that reconstruction into the next problem.

This seems to magnify, not diminish, the stakes with AI.

Generative AI changes this structure again, because the objective knowledge no longer merely waits to be retrieved.

6. Interactive objective knowledge

A search engine retrieves from an existing landscape. A generative model can recombine inherited knowledge around a problem that may not yet correspond to a document, course, or community.

You can therefore move through repeated cycles of problem, tentative solution, criticism, and successor problem while using a technological system as a rapid generator of conjectural material. The system can propose explanations, translate concepts, generate counterexamples, compare theories, critique arguments, produce code, expose hidden assumptions, and suggest rival formulations.

Yes, we're doing it now.

The important change is not that an artificial social mind has appeared. It is that objective knowledge has become interactive.

That interactivity makes the old distinction harder to keep in view. A book contains explanations but does not answer back. A calculator computes but does not converse. An LLM can produce explanations, arguments, code, criticism, plans, stories, and novel combinations. Ordinary language therefore slides easily into claims that AI knows, believes, understands, wants, or is smarter than humans.

Some of that may be harmless shorthand. The ontology remains unsettled.

Structure the problem so we can criticize it more clearly.

Three claims have to stay separate:

1. AI instantiates objective knowledge.
2. AI dynamically transforms objective knowledge.
3. AI is a knowing subject.

The first two can be true without the third having been explained.

The distinction becomes concrete during use. A system can greatly expand the candidate space of explanations without settling which problem matters, which criticism is decisive, or which successor problem should be pursued. You can reject the system's proposed trajectory and substitute another problem.

Current human-AI knowledge creation is therefore deeply coupled to individual problem selection. This does not prove that AI can never become an autonomous explainer. It identifies a difficult **problem of problems** within the present explanatory boundary.

The boundary is deeper than anthropomorphism. We still do not possess an explanation of how people create genuinely new explanatory knowledge. We can describe conjecture, criticism, problem selection, error elimination, recombination, analogy, and search. The mechanism by which you produce a genuinely new explanatory idea remains incompletely understood.

That incompleteness changes what can be claimed about machines. If explanatory knowledge creation itself is not yet explained, how can we assert that it exists in machines?

We can identify large stores of objective knowledge, linguistic and statistical structure, search, recombination, planning, problem-responsive generation, and criticism-like procedures. Whether those operations comprise the same universal epistemic process found in persons remains open.

Powerful output in adjacent areas therefore does not establish artificial explanatory knowledge. The limitation is not that AI is weak. The limitation is that the theory of knowledge creation is incomplete. Given the theory, the implementation would follow

quickly thereafter. And once explained, the epistemology we're applying here would be revealed as false.

"Superhuman" performance does not remove the gap. A telescope sees farther than a person. A calculator computes faster. A database stores more. A search engine retrieves across scales impossible for one mind. Generative AI extends this familiar pattern into tasks that were previously treated as close proxies for cognition.

A brief digression may be warranted. The superhuman capabilities of AI presently envelope fields that have been conventionally associated with super intelligence. A mathematician may look with despair on the progress of AI and conclude the singularity has arrived. It is still incumbent upon them to explain why. It is insufficient to point to the observations of mathematical proofs and induce from that proof of superhuman creativity. That would merely replicate the problem we've been discussing here. This problem of induction may inform the conjecture that AGI has arrived, we may conjecture AGI in the inductive direction, but it doesn't explain it.

A system can outperform an individual at a task without that performance, by itself, explaining a superior explanatory person. The artifact may do something astonishing. That performance still does not settle what kind of epistemic entity produced it, or how

Technology also makes the earlier political structure easier to scale. A censor once had to act repeatedly; a technical system can encode the censorship rule. A teacher once had to control access manually; a platform can automate restriction. A government once had to distribute instructions through officials; software can implement them instantly.

The same machinery that expands access to explanation can therefore expand constraint. Technology amplifies objective knowledge, including both explanatory and coercive knowledge. The relevant question is which knowledge has been encoded, by whom, and how that encoding changes the problem and action spaces available to particular persons.

The same technology can widen one person's field of possible criticism while narrowing another's. It can weaken one gatekeeper and create another. Once again, the explanatory unit is not "technology" acting as a subject. It is the relation between individual minds and the objective structures through which their possible encounters and actions are changed.

7. The recursive test

The argument can now be turned back on itself.

If the epistemology is universal, every apparent social mind must decompose to individuals. Psychology resolves into psychologists and objective disciplinary knowledge. A government resolves into officials, citizens, laws, records, technologies, and communications. A religion resolves into persons, texts, institutions, rituals, and inherited

explanations. A corporation resolves into employees, managers, procedures, contracts, software, and capital structures. Science resolves into scientists, instruments, papers, datasets, methods, and criticism.

AI must also decompose into technology, objective knowledge, training processes, system design, and interactions with individual people unless and until an independent explanation establishes a new epistemic subject.

If any one of these cases requires a special collective mind, then this universal account has failed. That is what makes recursion a test rather than an illustration.

The conventional frame can say that society believes, science knows, institutions decide, or AI understands because aggregate descriptions are useful. The epistemic frame keeps asking what the shorthand has compressed: Which individuals know what? Which objective structures preserve what? How is that knowledge communicated? Where is it reconstructed? Which person acts? Which structure changes the action space?

Some might call this intellectual dishonesty, or moving the goalposts, or denial of reality. We would say, reality is common sense, but epistemology is not, and universality provides the grounds.

Collective phenomena remain. What disappears is the need to add a collective subject merely because the pattern is stable or powerful. Society can be causally real, institutionally durable, and epistemically powerful without becoming a mind above the persons who compose and encounter it.

The epistemic frame can be compressed to three elements: individual knowers, objective knowledge, and fallible coordination between them. It's not a wisdom *of* the crowds, because crowds can't possess wisdom. It is a wisdom *between* the crowds, in the objective knowledge and interfaces through which individual minds exchange, preserve, criticize, route, and reconstruct knowledge.

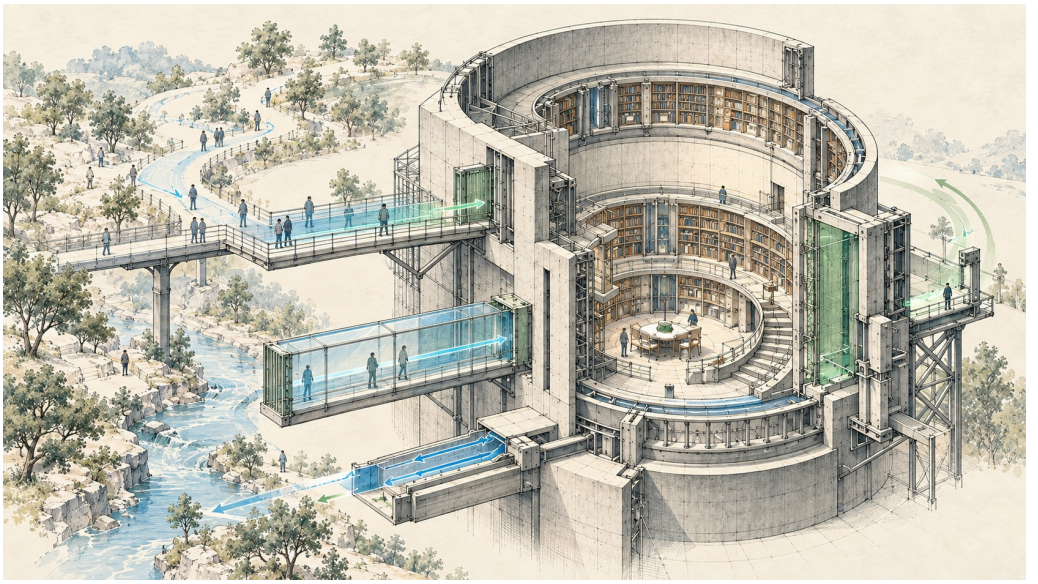
Every social structure in this account must be explainable in those terms. If one cannot be, the universality of epistemology has failed.

But survival produces a harder question. If society has no mind of its own, how does knowledge created by particular people become socially persistent? How is it reconstructed by people who did not create it, preserved across generations, transformed as it moves, and stabilized strongly enough to shape the problems of persons not yet born?

And if freedom or coercion changes which reconstructions remain possible, how does that alter the evolution of socially persistent knowledge?


One Reality, Many Worlds

V. The Popperian Heckle



You've been largely defensive. *You* should not be confused with genetic covariance, a psychometric category, a social class, an institutional role, or a collective abstraction. Genes matter. Psychological regularities matter. Social knowledge, institutions, and technology matter. But none of them is you.

Lots get positive then. What is the positive claim?

You is where objective knowledge can become a problem, where a problem can produce a conjecture, where a conjecture can be criticized, and where criticism can make a different problem possible. The schematic is:

$$P_1 \to TS \to EE \to P_2.$$

That doesn't sound hard. In fact, it sounds simplistic.

That's the problem with compressions: we're apt to lose the underlying explanation.

So what's hard about it?

The schematic $P_1 \to TS \to EE \to P_2$ is simple, the underlying epistemic schema is not. The schematic is useful because it describes the evolutionary logic of knowledge growth without pretending to explain its mechanism.

Like the familiar hard problem of consciousness, where describing neural and functional processes does not by itself explain qualia, this schematic leaves its own explanatory residue. We can name a creative problem, a creative tentative solution, a creative

criticism, and the successor problem that results, but the letters conceal almost everything important about how those states are instantiated and how the transformations between them occur. We do not yet understand, even in general, how you come to have this problem rather than another, create this explanation rather than one already available, or produce the criticism that changes an evaluation of the problem itself.

The underlying schema is therefore closer to an evolutionary state machine: it identifies a system (not a sequence) of epistemic configurations whose operation can be reconstructed historically, while leaving the generative content of those states and transitions largely unexplained.

Call that epistemic mystery **the Popperian heckle**. Richard Dawkins invented the term to describe pedantic philosophical arguments that are advanced to conceal manifest truth. Here we take that criticism seriously.

You're being defensive again.

If that epistemic cycle is not merely something we use to describe science or culture, but something that also occurs within your epistemic mind, what does it say about you. The question is no longer only what goes wrong when you are reduced to a category.

What becomes possible when you are treated as an open-ended creator of knowledge?

That question leads to freedom, but not first to freedom as the absence of physical constraint, an exemption from causality, or a political doctrine. A more primitive possibility appears earlier: your present knowledge does not contain all the knowledge you may later create.

A problem can become visible that is not visible now. A criticism can destroy an explanation that presently feels obvious. A distinction can appear that changes what you think the problem was. A future self can understand something the present self cannot yet think.

Your Popperian heckle is much more powerful than your selfish genes. Most of it remains unexplained. Some of it hides from criticism behind authorities of observation and manifest truth. That is the nature of universal explanations. They collide, they make us uncomfortable, even confrontational. But they don't care what we believe.

Your future self is still physically constrained, historically situated, biologically instantiated, and immersed in inherited objective knowledge. But no present description of what you know can also contain the explanations you have not yet created.

1. The Open Future

It offers us a different way to reflect on two old problems: If people move uniquely through evolving problem situations, how might free will and consciousness appear inside that movement?

The answer remains conjectural, as always. But the conjecture may reveal how much of freedom is not choice among already known alternatives, and how much of consciousness may not be a common container holding different contents. The epistemic individual may be more individual than that.

The explanation of *you* first appeared as something that had to be protected from explanatory substitution. But you are not merely what remains after false abstractions are removed. You are generative. You can encounter knowledge that already exists and produce knowledge that does not. You can inherit a problem and reformulate it, inherit an explanation and reject it, or inherit a value, institution, or category and discover that the problem it once organized has changed.

None of this implies infallibility. You also create errors. You misunderstand, become dogmatic, and construct explanations that make your own problem space smaller. The optimism lies elsewhere: if an error can be created, it can also become a problem. The present state is not final.

This gives freedom an epistemic structure. Freedom is often discussed as permission: permission to speak, leave, associate, or choose. Those freedoms matter because they affect what you can do. They also affect what you can come to know.

Your future depends partly on whether present explanations can still become problems. If an explanation cannot be questioned, a problem cannot be reformulated, rival knowledge cannot be encountered, or an unusable environment cannot be left, some epistemic transitions become unavailable. The same is true when you cannot try something whose value no authority can presently recognize.

> **Freedom preserves the possibility of movement into epistemic states that cannot be specified completely from the present one.**

This is broader than choosing from a menu. A menu contains known options. An epistemic future can contain alternatives that do not yet exist because criticism can change the problem itself. You may eventually reject every alternative you can presently name and create another.

Freedom is therefore not only a question of which available path may be chosen. It is also a question of whether the landscape of possible paths can continue to change.

That claim immediately creates a physical problem. The cycle

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2$$

compresses relations in the growth of knowledge. It does not imply that the brain contains a problem compartment, a tentative-solution compartment, an error-elimination compartment, and a successor-problem compartment. The symbols describe what states are doing in an epistemic process; they do not identify the mechanism that does it.

But the relations cannot float outside reality. If your state functions as a problem, that state must be physically instantiated somehow. The same is true of conjecture, criticism, and the transition by which one problem gives way to another. The implementation may be distributed, partly inexplicit, different across people, and badly aligned with the conceptual boundaries imposed by our notation.

The epistemology therefore does not give us a cognitive mechanism. It gives us a constraint:

> **if epistemic transformations occur, physically possible processes must instantiate them.**

The distinction matters because it lets epistemology generate questions about the mind without turning epistemic vocabulary into imaginary brain machinery.

Theory-laden observation gives one concrete clue about what such dependence can look like.

No jargon, please.

You do not first receive a neutral observation and only later attach an interpretation. What can be noticed already depends upon knowledge. A radiologist can see a structure in an image that another person cannot. A chess player sees a threat where a novice sees pieces. A mechanic hears a problem in an engine sound that another listener hears only as noise.

Abstract it.

$$T(O) \rightarrow O'.$$

The theory need not be an explicit sentence, and you may not know what objective knowledge is doing the work. The point is functional: prior knowledge changes what can become observable.

Now extend that structure cautiously. Let O_i denote, not a bare observation, but a physically instantiated epistemic state organized by the person's existing knowledge. Then we can write:

$$P_1(O_1) \rightarrow TS(O_2) \rightarrow EE(O_3) \rightarrow P_2(O_4).$$

The notation does not say that O_1 contains a little object called a problem. It says that this state functions as problematic for this person: something does not fit, attracts attention, or resists the current explanation. Likewise, O_2 is a state in which a conjectural

relation has become available, and O_3 is a state in which criticism has changed what can be retained.

The arrows remain schematic, and they also contain unexplained content. They tell us that transformation occurs, not how it occurs. Something moves you from confusion into a problem, from a problem toward a conjecture, from a conjecture toward criticism, and from criticism toward another problem. Sometimes that movement is explicit. Often it is not. The arrows may point in many directions at once. A sentence can feel wrong before the criticism can be stated; a solution can appear before its ancestry can be reconstructed; a criticism that seemed weak yesterday can become decisive today.

We do not yet possess an explanation of those transitions. But the ignorance is now more precise. The problem is not to find an uncaused mental event. It is to explain physically instantiated transformations whose epistemic role cannot be read off from a description of the present state alone.

That physical problem gives us a different entry into free will.

2. Freedom Inside a Mind

The traditional question often asks whether you could have chosen otherwise under the same physical conditions. That problem can remain where it is. The narrower epistemic question is:

> **what might free will look like inside a mind whose activity includes movement through changing explanatory states?**

Start with something ordinary. You can act in ways that change the conditions under which later thought occurs. You can redirect attention or ask a question, enter or leave an environment, consult an artifact, create one, or withhold one. These actions alter what you are likely to encounter next.

The plain relation comes first: you can change part of the environment that will later participate in changing you. Call that function **Choice**. The name does not refer to a proposed brain module or a little agent inside your mind. It compresses one relation: altering the epistemic conditions under which what happens next will occur.

An altered environment can change what becomes salient, which problem appears, which criticism becomes available, and which conjecture can be constructed. New knowledge can then alter later action.

Abstract it.

$$\Delta E \to \Delta K \to \Delta A \to \Delta W \to \Delta E',$$

where E is the person's epistemic environment, K their knowledge, A the actions that become available, and W the wider physical environment. The changed world becomes part of the next epistemic environment, and the loop continues.

Under this conjecture, free will may concern a causal process in which newly created explanations become new causes of later action. The explanation was not present before it was created. Once it exists, it can change what you notice, criticize, attempt, avoid, build, or seek.

The conjecture can therefore be stated cautiously:

> **Free will may be the evolving organization of epistemic openness: your capacity to participate in transformations that alter the conditions of your own future knowing and acting.**

This does not explain how explanatory novelty is produced. It changes the location of the problem. Instead of asking how freedom escapes causality or physical determinism, the inquiry asks how new knowledge can be physically created and then become causally relevant to what happens next.

Continuous evolving development makes the conjecture more individual. A capacity to alter epistemic conditions would not remain historically untouched, because every successful or failed attempt changes you before you attempt the next one.

People can therefore develop different architectures for performing related epistemic work. One person learns to decide by writing or speaking until a problem becomes explicit. Another builds rules or trained intuitions that reduce what must be reconsidered each time. Another learns to delay choice, seek disagreement, or interrupt the felt need to choose until the problem itself changes.

Abstract this choice architecture.

Call the developed form for individual i:

$$Choice_i(t).$$

The notation does not imply a discrete mechanism. It marks a possibility: the epistemic function may be shared while the developed architecture through which it is instantiated becomes highly individual.

This matters because the experience of choosing can vary. You may experience decision as forceful conscious authorship: *I chose this.* Training can change that experience. Meditation, for example, can make thoughts and intentions appear less like commands issued by an explicit chooser and more like events that become available to awareness. That phenomenological change does not by itself show that epistemic freedom has disappeared. The organization of choosing may have changed.

Injury can also alter forms of agency that had become central to one person's developed organization. Neurodiversity makes the observation universal. The point here is conceptual rather than clinical. Similar reports of choosing may arise from **different evolved architectures**, while different reports may arise from architectures performing related epistemic functions.

Free will, on this conjecture, would therefore not be one uniform inner object distributed across otherwise unique people. It would name related forms of epistemic movement whose mature organization becomes particular to a life.

That possibility creates the next problem. If choosing can become developmentally individual, perhaps conscious availability can too.

3. Consciousness and the Evolving Architecture

You embody and use far more knowledge than appears in your explicit consciousness. Biological knowledge, perceptual expectations, motor skills, language, memories, learned salience, emotional associations, habits, unfinished problems, and inexplicit criticisms can all affect what happens without first appearing as explicit objects of awareness.

Some of this becomes conscious. Much does not. Consciousness therefore cannot simply be identical to all the knowledge operating you.

The narrower epistemic question is:

> **what changes when some part of your larger epistemic organization becomes available in a form that can participate differently in criticism, comparison, communication, or reconstruction?**

A pain can become a problem rather than only a disturbance. A memory can become something that is compared with evidence rather than merely something that biases action. An intention can become something that is inspected and criticized before it is acted upon. Inexplicit unease can become a sentence: *something does not fit.*

Perhaps consciousness includes transformations through which parts of a much larger epistemic organization become available for forms of error correction that would otherwise be unavailable. That is a conjecture about what conscious availability may do in the epistemic cycle.

Development now returns. Shared biological capacities do not imply identical mature or static epistemic organization. People acquire different languages and distinctions, practice different forms of attention, inherit different social explanations, and turn different skills into inexplicit competence. Those histories change what later becomes salient and what criticism can reach.

Why assume that conscious organization is exempt from that history or that the process ever stops?

Perhaps people do not merely have different contents inside one common conscious processor. Perhaps some of the organization through which content becomes consciously available also develops. One person may rely heavily on inner speech while another reports little of it. One may organize thought visually or spatially while another relies more on bodily salience or explicit conceptual chains. One may experience agency as a

strong internal act while another experiences an intention as something that seems to arrive.

These differences support a narrower conjecture:

> **evolving consciousness may be a shared biological possibility expressed through individually developing epistemic architectures.**

If so, the familiar hard problem remains. Nothing here explains why physical processes are accompanied by subjective experience at all. But another problem becomes visible beside it: Are you carrying the implicit assumption that the conscious organization whose subjectivity you are trying to explain is one uniform object across evolving persons?

The opacity of another mind may therefore arise not only because experience is private. It may also arise because the structures through which experience becomes epistemically available have been shaped by a history no other person shares. This does not solve the mystery. It makes one part of the mystery more local and more individual.

The free will and consciousness conjectures meet at one correction. Freedom cannot require complete conscious control over the causes of thought. You do not explicitly choose every problem that becomes interesting, command every useful memory to appear, or issue an internal instruction that generates each conjecture. You can know that something is wrong before knowing what, and the next thought can surprise you.

If freedom required a conscious executive directing every epistemically relevant transition, the epistemic mind would appear mostly unfree. But the conjecture developed here requires less: *You* can participate in changing the conditions of future thought without controlling every transformation that follows.

You can learn, externalize a problem, practice a skill, seek criticism, build a tool, alter a relationship, or change what you expose yourself to. Or you could choose not to. Each intervention modifies part of the environment in which later cognition occurs. The altered environment changes what you encounter; what you encounter changes what becomes salient; what becomes salient can change your next problem.

The mind changes the environment. The changed environment returns to the mind. You change.

Epistemic freedom may therefore operate recursively across explicit and inexplicit processes. You need not be transparent to yourself to remain the locus of epistemic development. Self-opacity and agency are not opposites.

That result is interesting because your environment is not made only of physical objects. It is already full of explanations created by other people.

4. When Social Knowledge Enters

You enter a world already filled with knowledge embodied in language, family explanations, institutions, media, laws, moral concepts, role descriptions, categories, and technologies. This is objective knowledge in the earlier sense: knowledge that can exist outside any one person's present awareness and be encountered, reconstructed, criticized, or reused.

You do not receive this knowledge as a neutral object. You reconstruct it through an existing epistemic organization. Once reconstructed, social knowledge can participate in your own problem situation. It can open problems, but it can also close them.

Consider explanations such as *this is what I am*, *this failure reveals something permanent*, or *this problem has no alternative solution*. Each can become internally causal without subsuming you. An inherited explanation can affect what you notice, which alternatives you consider, what criticism feels available, and which successor problems can even be formulated.

This gives social knowledge a concrete route into individual life. The damage need not be a mysterious social force acting from outside. An explanation enters your epistemic environment and reorganizes what movement appears possible from inside it. The same mechanism also explains how social knowledge can heal: a new explanation can reopen the landscape.

Consider someone who explains their situation this way:

> I am defective because I lack conventional social success.

The explanation locates the problem in a fixed property of the person. It also constrains what can follow. If the cause is an essential defect, many possible next questions disappear before they are asked.

Now consider a different explanation:

> I have not yet solved the problem of finding relationships in which the things that matter to me can be shared.

The outward condition may be identical. The epistemic condition is not. The second explanation creates different successor problems: where are such people, what forms of communication could make connection possible, which inherited expectations misdescribe the problem, and what kind of community might not yet exist?

A better explanation need not remove loneliness. It changes what loneliness can lead to. You occupy a different problem space because different successor problems have become available.

Objective social knowledge can enter your mind and alter the sequence of epistemic transitions that follows. You can inherit an explanation that closes movement. You can also encounter one that makes movement possible again.

Once freedom is conceptualized this way, flourishing need not produce greater psychological convergence. The opposite can occur because expanding objective knowledge gives people access to more specialized epistemic environments: unusual disciplines and technical subcultures, distant communities and languages, new artistic forms and explanatory traditions.

The reachable epistemic landscape becomes larger, but individuals need not distribute themselves through it in the same way. Greater opportunity can therefore generate greater divergence. You may become increasingly unlike the people nearest you in interests, knowledge, language, explanatory commitments, skills, and problems.

From the perspective of an aggregate, divergence may look like fragmentation. From your perspective, it may be flourishing. Divergence itself settles neither interpretation. You can diverge into error, isolation, dogma, or obsession.

The narrower claim is:

> **distance from the mean cannot tell us whether an epistemic trajectory is flourishing or failing.**

That requires criticism of the trajectory itself. The same conclusion now reaches inward. If epistemic histories diverge, the architectures produced through those histories may diverge too. Different people may not merely think different things. They may differ in how problems become salient, how choice is organized, how criticism is experienced, and how parts of the mind become consciously available.

You become hardest to infer from the population precisely where your history has done the most work.

Divergence then creates a practical problem for anyone responsible for another person. A parent may watch a child enter technological environments the parent does not understand. The child develops interests, friendships, language, and problems that no longer overlap neatly with the parent's world. The uncertainty is real, and so is the loss of legibility.

The epistemic error begins when one explanation is substituted for another:

> I no longer understand the child's world

becomes:

> the child's world is harmful.

Sometimes it is harmful. Sometimes restriction is necessary. But unfamiliarity and harm are different claims. Restoring an environment the parent recognizes can restore parental legibility without necessarily improving the child's epistemic condition.

The same structure can appear in schools, workplaces, institutions, and governments. When another person's trajectory becomes difficult to interpret, control becomes attractive because control reduces variation. Reduced variation makes the system easier for the controller to see. That is an observable effect. It does not by itself tell us what happened to the controlled person's problem space.

Legibility and flourishing can therefore come apart. Once they do, the problem is no longer merely interpersonal. It becomes a problem about coordination, authority, and the preservation of criticism.

5. Freedom, Legibility, and Epistemic Closure

Authority can solve real coordination problems. It can act quickly, concentrate expertise, set standards, and prevent local actions from imposing unacceptable costs on others. Nothing in epistemology eliminates those functions.

The problem begins when responsibility for a problem is mistaken for possession of its uniquely correct solution. The epistemic cycle is:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2.$$

An authority can instead produce:

$$P_1 \rightarrow TS_A \rightarrow \text{control}.$$

The danger is not authority as such. It is closure. A problem definition becomes mandatory, alternative conjectures disappear, criticism becomes disobedience, and a successor problem is reinterpreted as refusal to accept the solution.

The immediate system may become easier to coordinate while becoming less able to discover errors in the explanation governing that coordination. Coordination and criticizability are therefore not opposites. A coordinated response can still preserve routes through which dissent, revision, exit, competing experiments, or replacement can occur.

It is not only whether anyone is in charge. It is whether the explanation governing the intervention can itself still become a problem.

Pressure sharpens the tension. The more serious a problem appears, the stronger the temptation to close the epistemic cycle. High stakes make variation look dangerous, delay irresponsible, and criticism obstructive. But urgency does not transform a tentative solution into infallible knowledge.

That creates an asymmetry between effects that exist after an intervention and alternatives prevented from existing. A restriction may reduce a measured exposure, a rule may increase compliance, or a ban may stop a particular action. Those effects can often be observed. The suppressed alternative is different.

If you are prevented from pursuing a problem, the conjecture you would have created does not exist. If an institution suppresses a criticism, the successor problem that criticism would have produced may never become visible. If a child is prevented from entering an unfamiliar domain, a later skill or relationship that might have emerged cannot appear in the data.

Compress it.

Coercion can produce:

$$\text{known compliance} + \text{unknown epistemic loss}.$$

The second term cannot be inventoried in advance because it refers partly to knowledge that was never created. That does not imply that every restriction is wrong or that emergencies require decentralization. Sometimes rapid coordination is indispensable.

The narrower point is epistemic:

> **severity increases the cost of error in the imposed solution as well as the cost of error in failing to act.**

A system can therefore gain coordination while losing some of the mechanisms through which errors in that coordination would have been discovered. The loss is hardest to measure when the prevented future never becomes actual.

6. The Expanding Epistemic Environment

The same open-future argument applies to physical constraints. Matter, energy, time, and infrastructure are real. You cannot explain scarcity away. But the practical meaning of a physical constraint depends partly on knowledge.

Oil existed before people knew how to use it as an energy resource. Uranium existed before people knew how to release useful energy from it. Silicon existed before people knew how to build computational systems from it. In each case, the material world set constraints while knowledge changed which possibilities inside those constraints became reachable.

Abstractly:

$$\text{present constraint} \neq \text{permanent possibility boundary}.$$

The climate and energy demands associated with artificial intelligence provide an immediate example. AI infrastructure consumes real electricity and materials, creating real problems. But a description based only on current energy technology cannot contain every future way of responding to those problems. Compute demand can create pressure

for better hardware and software efficiency, different cooling and storage, changes in generation and transmission, or technologies not yet available.

Innovation may fail, arrive too slowly, or create new problems. Nothing in epistemic openness guarantees rescue. The narrower claim is that the future resource landscape depends partly on explanations that do not yet exist. Present scarcity is evidence about present constraints. It is not a complete inventory of future practical possibility.

Artificial intelligence matters here for another reason. It can change not only the resources available to society, but the problems you can reach. You can use it to cross unfamiliar technical vocabularies, generate and test candidate models or code, and revisit bodies of objective knowledge at scales unavailable to unaided memory. You can use it to cross unfamiliar technical vocabularies, generate and test candidate models or code, and revisit bodies of objective knowledge at scales unavailable to unaided memory.

We're doing it now, for better or worse.

The relevant consequence here is immediate: the technology changes your epistemic environment.

You can reach problems you previously could not practically enter and encounter rival explanations faster. You can construct tentative solutions with tools you did not possess and expose a conjecture to forms of criticism that were previously too costly or inaccessible. That enlarged reach can produce error, but it can also expose error.

Abstract it.

The technology therefore enters the reciprocal process:

$$\text{objective knowledge} \rightarrow \text{individual reconstruction}$$
$$\rightarrow \text{criticism and creation} \rightarrow \text{new objective knowledge.}$$

You remain the locus at which inherited knowledge becomes a live problem and at which a generated output is accepted, rejected, misunderstood, corrected, or transformed. AI changes the reachable landscape. It does not displace the epistemic mind.

This reciprocity prevents the argument from becoming individual against society. You do not create from nothing. You begin inside objective knowledge created by others: language, mathematics, crafts, moral vocabularies, scientific explanations, stories, legal concepts, technologies, and ways of naming experience.

You may suffer partly because the explanations available to you are poor. A new distinction can then reorganize the problem. A failure can become a learning problem rather than a verdict on worth. A relationship can become a coordination problem rather than proof of defect. A diagnosis can become one useful description rather than a complete identity.

The physical world may not change immediately. Your problem space does. That is one way social knowledge can heal.

The recursion moves outward.

Changed, you can criticize inherited knowledge and externalize something new: a distinction, tool, explanation, institution, method, or previously unformulated problem. Social knowledge can therefore be changed by precisely the kind of person that existing categories do not yet describe well.

The dissenter, eccentric, difficult student, lonely outsider, unconventional researcher, obsessive engineer, or misunderstood artist may ask why something has to be the way it is. Most such questions will not produce important objective knowledge by conventional measures. That does not negate the opportunity. An important criticism cannot always be identified before it is created.

You are therefore not outside social knowledge. You are a place where social knowledge can change.

This is also why more freedom need not make people more alike. More objective knowledge need not produce one common epistemic landscape. Better communication need not eliminate deep disagreement, and greater technological reach can increase differentiation.

You can find collaborators around an interest nobody nearby understands. A child can encounter a discipline unknown to the family. A researcher can combine fields whose institutional boundaries previously kept them apart. These developments can produce fragmentation, but they can also produce flourishing.

Uniformity can arise from shared good explanations or from restricted possibility. Divergence can arise from error or from expanded possibility. Neither convergence nor divergence therefore settles the epistemic question.

The trajectory has to be criticized at the level of the problems, explanations, errors, and consequences actually involved. No population metric can automatically tell us whether your epistemic path is flourishing merely because it is common or uncommon. Freedom preserves variation partly because future knowledge cannot be identified in advance by conformity to present knowledge.

That returns us to the boundary no abstraction can cross for us: another person.

7. Another Person

We cannot leave our own minds, directly occupy another person's epistemic state, or inspect their conscious organization from inside. We encounter speech, action, artifacts, silence, expression, and history. From that surface we construct an explanation.

Understanding another person is conjectural. The explanatory questions become concrete: what problem are they trying to solve? Which distinctions matter to them, and what history made those distinctions salient? What alternatives do they think are available? What knowledge or transition am I failing to see?

Then the other person can criticize our model.

No. That is not me.

An error becomes visible, and the model changes. That is progress. The goal is not fusion. Two minds do not need to become the same to understand one another better. The distance can remain while the bridge improves.

That distance can itself become epistemically useful. A person whose world overlaps heavily with ours may leave many of our assumptions invisible. A person who is very different can expose them. They may reject a distinction we never thought to question, notice a problem we cannot see, or misunderstand us in a way that reveals how poorly we explained ourselves.

Difference therefore has epistemic value even when agreement never arrives. It creates surprise, and surprise can create a problem. The stranger can reveal the invisible boundaries of our own landscape. Interpersonal distance is then more than a communication obstacle. It can become an engine of error discovery.

If choice and conscious availability are partly shaped by individual epistemic development, the distance may run deeper than disagreement over beliefs. People may differ in some of the structures through which salience, choosing, criticism, and conscious availability are organized.

That possibility weakens the assumption that your own internal form can serve as a universal template for another person. It also completes the initial movement of this series. You first had to be protected from explanatory substitution. You then became a positive epistemic object: a physically instantiated, socially embedded, historically developed locus at which inherited objective knowledge can become problematic and new knowledge can appear.

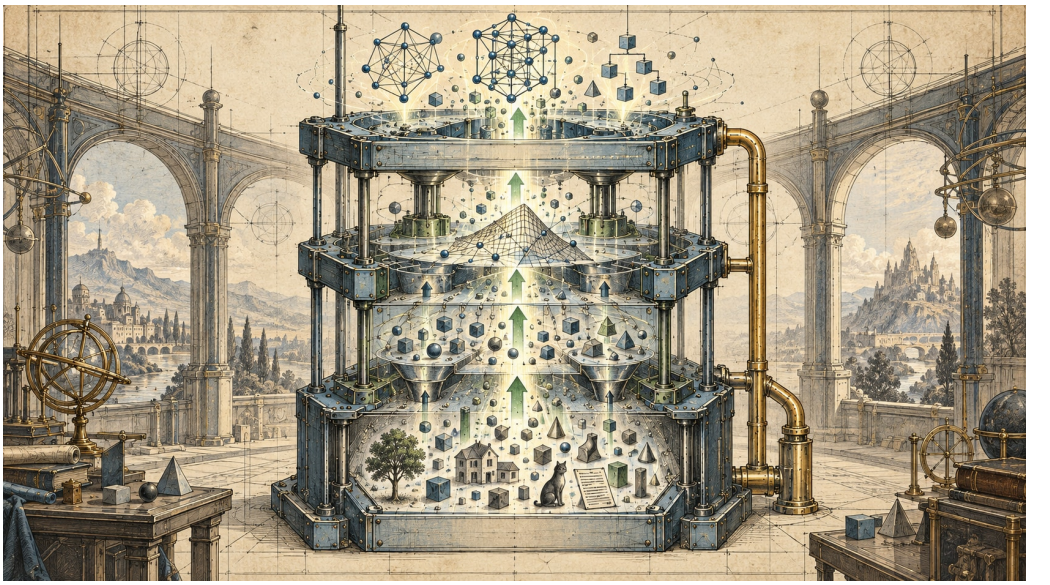
The same account that makes you generative also makes another individual difficult to compress.

Every person inherits different objective knowledge, encounters different problems, develops different inexplicit expectations, and follows different epistemic transitions. If evolving architectures of choice and conscious availability also become increasingly individual, the distance between minds may be much larger than ordinary social categories suggest.

There is one reality. But there may be billions of deeply path-dependent ways of experiencing and explaining it.

How can persons who share one reality come to inhabit epistemic worlds that are so profoundly different, and what does that imply about the abstractions through which each person experiences both reality and everyone else?

VI. The Black Rabbit Hole



You cannot leave the locus of your own mind, but you can still move toward another person's world through conjecture, communication, and criticism. You can ask what problem the other person is trying to solve, construct a tentative model of their world, allow them to criticize it, and sometimes return with an error in your own understanding exposed.

If that's true, the distance between minds is enormous.

People can share a language, a country, a family, even decades of common history, yet inhabit profoundly different experienced worlds. Religious beliefs, conspiratorial explanations, and political convictions can remain resilient in the face of contrary observations. At the same time, you readily experience your own inner life as deep and complicated while strangers, opponents, and out-groups appear comparatively simple.

There is only one reality.

The explanation developed here does not require many realities. There is one reality. But you do not encounter it whole or without mediation. Perception and thought select, organize, retain, and omit. They make some distinctions available, some problems visible, some expectations natural, and others difficult even to formulate.

Call the resulting structures **abstractions**. They are not intellectual decorations placed on top of an otherwise direct encounter with the world. Mediation begins in

perception itself and continues through inexplicit expectation, learned salience, concepts, language, memory, social knowledge, theories, institutions, and explicit explanation.

These abstractions are real because they have causal effects. They can be true, false, partial, useful, misleading, incoherent, resilient, or mutually incompatible, yet still alter what you notice, remember, trust, fear, attempt, and criticize. Most are not available to introspection as explicit propositions. They become visible when some encounter, contradiction, criticism, or problem forces them into view.

That is stronger than saying only that abstractions are real. It says that experienced reality is mediated through them, including at the level of perception.

Relativism is false.

The claim is not relativist. The abstractions differ, but the reality constraining them does not multiply with the observers. Every abstraction remains fallible within the same reality.

The universal epistemology used here must submit to the same rule. The claim that persons construct evolving worlds of knowledge is itself a constructed explanation. So is evolutionary epistemology. So is the distinction between reality and representation. None becomes exempt from criticism because it appears superficially foundational. Realism does not grant certainty. It gives criticism something beyond the abstraction to answer to.

Compare that idea.

Let R represent mind-independent reality and M_i the organized world through which individual i encounters it:

$$M_i \neq R.$$

But M_i is not unreal. It is instantiated in the person and changes what that person can do. A representation can therefore be false about what it represents while remaining real as a cause. A conspiracy theory can be false about its referent and still organize suspicion, memory, and action. A religious doctrine can contain false factual claims while still connecting morality, identity, suffering, family, ritual, community, and purpose. A political abstraction can misdescribe institutions while still shaping alliances, moral judgment, and perceived threats.

The distinction matters because the same error can occur at many levels: a psychometric construct becomes personality, a population category becomes the person, or a social statistic becomes a causal mechanism. In each case a representation stops being treated as a selective structure and begins to stand in for the thing represented. Universal epistemology would repeat the error if it were treated as direct access to reality rather than as a tentative explanation of how access works.

Mediation begins before explicit belief. We do not first receive raw reality and then voluntarily place concepts over it. A face does not appear as an uninterpreted

distribution of wavelengths and only later become a face. A familiar voice, a threat, and another person already arrive through perceptual systems shaped by biological history and individual development.

Develop the abstraction.

The sequence is closer to:

$$R \to \text{perceptual selection} \to \text{inexplicit abstraction}$$
$$\to \text{salience} \to P \to \text{possible explicit conjecture.}$$

Reality constrains this process continuously. A perceptual system cannot freely invent a cliff where no relevant physical structure exists and remain indefinitely successful in acting on that invention. But constraint is not unmediated access. The experienced world is already selected and organized.

Most of that organization remains inexplicit. It operates through learned salience and expectation, embodied skill and tacit causal structure, and perceptual categories bound up with trust and emotional orientation. You normally do not think, *I am applying abstraction A, which depends on assumptions B, C, and D.* You experience: *this is what is happening.*

And then reality kicks back.

A failure can change that. A familiar solution stops working. A prediction is violated. Two trusted sources conflict. A relationship produces an unexpected response. What had functioned invisibly can now become a problem. Criticism therefore does more than test explicit statements. It can expose the structure through which the statement first appeared obvious.

That general claim becomes concrete when we look at a representation designed precisely to make different people comparable. Population psychology often places persons into a common coordinate system:

$$\text{person} \to \text{standardized observation} \to \text{common dimension.}$$

The transformation is useful because comparison requires commensurability. It is also a compression: only differences expressible in the chosen measurement space survive. The rest of your organized history is outside the representation by design.

That does not make a trait score, population distribution, or longitudinal model false or useless. It tells us what kind of knowledge the representation contains. A trait score can locate a person relative to others on a defined dimension. A population model can describe regularities in a sample. A longitudinal model can describe change over time. None of these, merely by existing, contains the developmental history that produced your present organization of problems, expectations, distinctions, skills, trust, and salience.

The epistemic difficulty appears when the common representation is applied back to you. Population regularities and individual histories are different objects of explanation. Your membership in a population does not by itself supply the causal bridge from a group-level relation to your development. That bridge has to be discovered rather than assumed.

This creates a conjecture that can fail. If increasingly fine-grained explanation converges on the same stable causal architecture across persons, then the strong individuality thesis weakens. If similar surface scores and behaviors repeatedly resolve into different problem histories, causal pathways, and person-specific dynamics, the thesis gains corroboration.

The issue cannot be settled by declaring individuality in advance.

The common representation has now done something more important than merely classify. By seeing exactly what it preserves, we can see exactly what it leaves open. The problem shifts from how people vary inside a shared measurement space to what kind of thing the singular person must be if every common representation compresses part of the history that produced them.

Explain.

The evolutionary story begins with biological knowledge. Genes embody inherited solutions to recurring problems of organism construction and survival. They participate in constructing sensory, memory, motor, affective, attentional, and learning systems, including a brain capable of language and abstraction. This shared biological knowledge gets you started without specifying the semantic content of the person you will become. The future world of meanings, problems, and explanations is not present in miniature inside the starting biology.

Development therefore cannot be described simply as the accumulation of inputs. At any moment you possess some existing knowledge K_t. That knowledge makes certain problems visible and others invisible. You construct tentative solutions. Those solutions encounter criticism from reality, action, other people, contradiction, surprise, and later knowledge. Some are retained, some revised, some abandoned. The retained structure changes the starting point for what happens next:

$$K_t \rightarrow P_t \rightarrow TS_t \rightarrow EE_t \rightarrow K_{t+1}.$$

The crucial relation is not merely that knowledge changes. It is that K_{t+1} changes which future problems can exist for you. Earlier knowledge alters the environment in which later knowledge develops. That recursive dependence is the mechanism of epistemic individuality.

Give me an example.

Two children can begin from biological starting points that are almost indistinguishable for ordinary human purposes and still construct very different regions of experience.

Both hear music. One becomes curious about why a melody sounds unfinished. The curiosity becomes a problem; an explanation about harmony offers a tentative solution; experiment reveals where the explanation works and fails; a new distinction becomes audible. Years later the person hears tension, phrasing, modulation, and structure that the other child does not notice. A region of the person's world has been constructed.

The second child encounters a programming puzzle. A distinction between state and process becomes salient. Debugging becomes a skill, then a style of questioning that later enters mathematics, work, perhaps even interpersonal reasoning. Neither trajectory was contained in the initial biological difference. Each emerged because earlier knowledge altered later problem selection. After twenty years, the two people can stand in the same physical room and encounter different problems in the same event.

That is the relevant meaning of **path dependence**. It does not mean that tiny differences must explode or that trajectories cannot converge. It means that retained knowledge changes what can happen epistemically next.

Abstract that path.

$$\Delta K_t \to \Delta P_{t+1} \to \Delta TS_{t+1} \to \Delta K_{t+1}.$$

A difference in knowledge changes the next problem; a different problem invites a different conjecture; a different conjecture encounters different criticism; the retained result changes the next epistemic environment. Over a lifetime, the sequence compounds.

Human abstraction makes the resulting space open-ended in a further sense. People can construct mathematics, moral and legal systems, and programming languages whose objects and rules need not have existed before those abstractions were created.

More important than any list is the recursion: abstractions can operate on abstractions.

A programmer can create a language used to build another language. You can form a theory about another person's theory about an institution whose rules were themselves produced from abstract legal concepts. The next problem need not be available inside the conceptual world that existed before the new abstraction was created.

There is therefore no known fixed upper bound on the possible organization of an epistemic mind. This gives precise content to the metaphor that you can become **a universe unto yourself**: not a separate reality, but a vast, historically evolved attempt to explain and navigate one reality.

Most of that world is not explicit, is not explained.

It includes skills that cannot be fully verbalized, patterns of salience, anticipations, emotional associations, embodied expectations, implicit distinctions, memories whose significance is only partly available, and causal expectations revealed only when violated.

You can be unable to explain why something feels obvious even when that feeling has a developmental history. The explanatory structure is larger than introspection.

Criticism can make part of it visible. Ask why a source is trustworthy and an inherited institutional model may appear. Ask why a behavior counts as disrespect and a tacit moral category can become explicit. Ask what observation would change your mind and you may discover that the theory had no imagined failure condition. The criticism tests the explicit proposition, but it can also expose the abstraction that made the proposition feel immediate.

When such an abstraction becomes deeply embedded, you may stop experiencing it as a model at all. You do not experience *my model says X*. You experience *obviously X*.

Call this **manifest truth**: an abstraction whose mediating role has disappeared from experience and come to represent reality.

$$\text{inexplicit abstraction} \rightarrow \text{experienced immediacy} \rightarrow \text{manifest truth}.$$

A manifest conviction need not be false. The error lies in treating immediacy as evidence that mediation has vanished. The abstraction can be hardest to notice precisely when it operates most fluently.

This also explains why large beliefs can be resilient. A belief is rarely an isolated sentence. It may be connected to family and identity, source credibility and institutional trust, or suffering and purpose. The particular network differs by belief, but the causal consequence is the same: criticism aimed at one visible proposition can encounter a much larger structure.

A sufficiently flexible structure can then route around anomalies. An observation may be reinterpreted, rejected because of its source, or absorbed into a larger story. This is not peculiar to conspiracy thinking. Any explanatory structure flexible enough to accommodate every possible result can protect itself. The important distinction is between resilience produced by explanatory depth and resilience produced by unlimited flexibility. Only the latter makes failure conditions disappear.

Religious belief, conspiratorial thinking, and political polarization can therefore exhibit similar structural features without becoming equivalent in truth or content. A religious belief can be embedded in moral, familial, existential, and communal meaning. A conspiratorial model can interpret criticism through the theory itself. Political disagreement can place the same observation inside different networks of trust, history, causation, identity, and moral salience. In each case the disagreement may extend far beyond the visible proposition because the proposition is intelligible only inside a larger epistemic architecture.

That is why adding facts can fail to resolve a disagreement. The facts do not enter an empty mind. They enter an already vast evolved world.

The same mechanism becomes more consequential when a manifest truth is joined to authority. Authority need not operate only by issuing commands. It can operate by treating certain propositions as prior to criticism: *any reasonable person can see this*; *decent people know this*; *this is simply reality.*

Abstract the transition.

manifest truth → dissent reclassified as defect.

A rival explanation is no longer encountered as a rival explanation. The dissenter is reclassified as irrational, immoral, ignorant, dangerous, or unfit to participate. The abstraction is then protected from the criticism that might have exposed its hidden assumptions.

The same ontology explains a quieter interpersonal error. Your inner world appears rich because you inhabit it. Another person's appears thin because you encounter only a surface. You have direct access to your memories, unfinished problems, and inexplicit expectations. You encounter the other person through speech, behavior, and occasional disclosure.

Abstract the comparison.

thin observable surface → thin model of other

but not:

thin mind.

The sparse thing is usually your abstraction of the other person.

Social labels make the compression easy to see. People close to you are permitted complexity because you know their exceptions, histories, contradictions, private failures, peculiar interests, and changing moods. An outsider may be compressed to *conservative*, *religious*, or *conspiracy theorist.* A label can serve a narrow descriptive purpose. It becomes an epistemic error when the compression is mistaken for the person.

Biological differences do not repair that inference. People differ in processing speed, memory, and sensory or attentional capacities. Those differences can matter. Ordinary human minds nevertheless display open-ended explanatory generality, and these implementation differences do not by themselves establish that one person's epistemic mind is categorically richer than another's. A slower processor is not thereby a shallower explainer, and weaker recall does not by itself imply a barren internal world. Writing, software, and AI can also externalize or extend some capacities, further loosening the link between a particular implementation difference and the possible depth of a person's explanations.

Nor does the epistemic process occur only during science or philosophy. A conversation creates expectations; a disappointment creates a problem; a hobby generates

new distinctions. Each retained solution can change what becomes salient next. Over years, ordinary life produces an enormous developmental history even when little of that history is ever rendered as explicit theory.

This reverses the usual evidential burden concerning strangers. A person who has spent decades perceiving, learning language, solving practical problems, forming relationships, remembering, imagining, failing, and revising expectations is predicted to possess substantial hidden structure. Depth is not truth, and complexity is not explanatory quality. A vast epistemic mind can contain enormous error. But lack of access to another person's depth is not evidence that the depth is absent.

Reconsider the psychological case.

Now the objective knowledge supplied by the psychological case can be reconsidered without asking it to do more than it was built to do. A population model can be accurate at the level for which it was constructed. A trait score can locate a person on a common dimension. A group-level regularity can describe a sample. Such knowledge can support comparison, prediction, and further inquiry. What it does not contain merely by being objective is the person's full causal and developmental architecture.

The person-level account changes the explanatory demand. The population abstraction need not explain the whole person, and its failure to do so does not make it useless. The sharper question is what explanatory work the representation performs and where that work ends.

The individuality conjecture remains open at that boundary. If increasing explanatory resolution repeatedly reveals different path-dependent histories and causal organization beneath similar surface scores or behaviors, the thesis survives.

What??

Biological knowledge evolves organisms capable of perception, memory, abstraction, criticism, and further knowledge creation. From there, retained solutions alter future problem situations; new abstractions alter salience; criticism can expose structures that had previously operated invisibly.

This is how people can share one reality while inhabiting profoundly different experienced worlds. It is also why disagreement can remain deep even when language, country, workplace, family, or decades of history are shared. Visible overlap does not reveal the full networks of memory, trust, morality, causation, identity, and explanation that give an observation its meaning.

The asymmetry of other minds follows from the same cause. Your own world is available to you in immense detail. Another person appears through a compressed surface. The temptation is to mistake your model of that person for the person themselves. *My inner world is rich; yours is barren* then looks like a discovery about the other person when it is often a discovery about the poverty of your model.

None of this weakens realism.

It makes realism necessary to the argument. If there were only private realities, criticism would lose its discriminating force. Because there is one shared reality, every epistemic mind remains answerable to something beyond itself. People can be wrong. Belief systems can be wrong. Institutions and cultures can be wrong. Universal epistemology can be wrong.

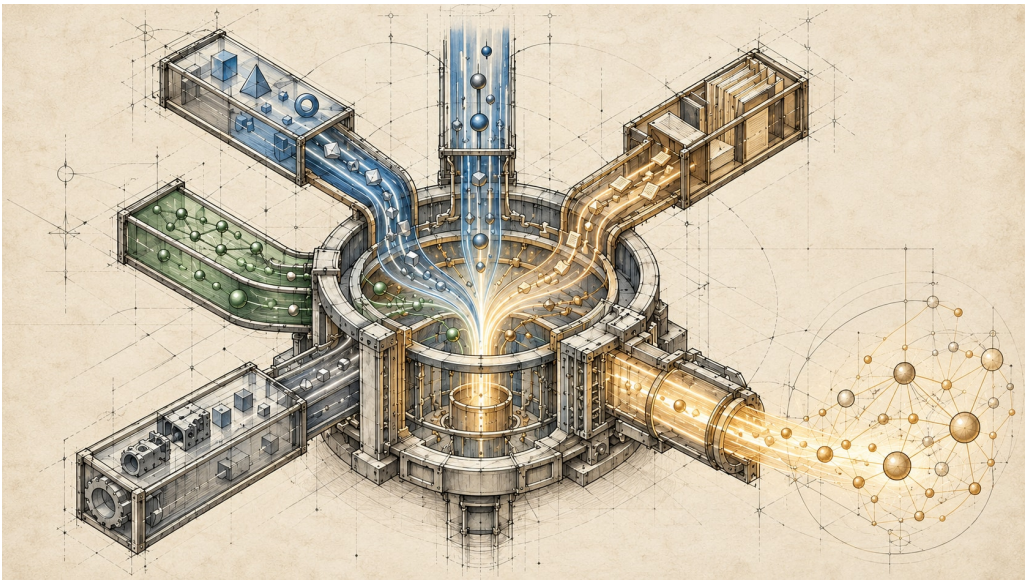
The process therefore remains:

$$\text{construct} \rightarrow \text{criticize} \rightarrow \text{reconstruct}.$$

Reality is the ultimate constraint, and no individual possesses it whole. The epistemic landscape is vaster than you imagine, much of it is inexplicit even to the person who inhabits it, and the map you carry of another person is much smaller than the territory.

If every encounter with another person is mediated through your own evolving abstractions, and if the abstractions most in need of criticism may be precisely those experienced as reality itself, then the problem is not merely whether epistemic distance can be crossed. It is how criticism can reveal the invisible abstractions on both sides of the crossing.

VII. The Emissaries and Their Master



If individual minds diverge so radically, what prevents them from dissolving into arbitrary private worlds?

There is one reality.

Explanations may begin in different domains, at different levels, and in different minds, but explanations of the same world cannot remain indefinitely incompatible. Physics, computation, evolution, epistemology, psychology, morality, institutions, and ordinary experience do not inhabit separate realities. Their explanations must remain mutually possible.

When independently developed explanations fit together well enough that no unresolved contradiction presently forces reconstruction, explanatory knowledge has reached a provisional equilibrium. That state will be called **consilience**.

1. Consilience as explanatory equilibrium

It is not proof, consensus, psychological comfort, or accumulated confirmation. Because knowledge is fallible, the equilibrium is temporary. A new criticism can disturb it. A contradiction can become a problem. A new explanation can reorganize the whole.

Consilience is therefore not the end of the epistemic cycle. It is one of its resting states:

$$\text{equilibrium} \rightarrow P_1 \rightarrow TS \rightarrow EE \rightarrow \text{new equilibrium}.$$

The letters are doing too much work here, but the abstraction remains: moments of epistemic equilibrium unsettled by states of problems, tentative solutions, and error elimination.

The question is how that equilibrium operates across many domains when criticism and reconstruction still occur in you. Your biological processing capacity is finite.

Different explanations can begin independently without being independent of reality.

Physics constrains what transformations can occur. Computation constrains what transformations can be represented and executed. Evolution explains how adaptive structures can arise and persist without a conscious designer. Epistemology explains how you can create and improve explanations within that world.

These explanations do not reduce to one another. They answer different problems. But they cannot require mutually impossible worlds.

If one explanation requires something that another good explanation says cannot exist, the contradiction is not merely disciplinary. It becomes a problem for the explanatory system as a whole.

Abstract it.

Consilience can be represented as a dynamic relation among explanatory structures:

$$K_1 + K_2 + \cdots + K_n \rightarrow \text{provisional compatibility}.$$

The word *provisional* matters. Every explanation remains criticizable. A new observation, criticism, or explanation can expose a contradiction that had not previously been visible and force reconstruction.

That follows from realism. If there is one reality, then every valid explanation concerns some part or aspect of the same world. A theory of mind cannot require processes forbidden by physics. A theory of knowledge cannot depend on transformations that cannot in principle occur. An account of human individuality cannot ignore biological development. Explanations of actual persons can, in turn, expose problems in psychology or social theory. The relation is reciprocal constraint, not a hierarchy of disciplines.

Evolutionary epistemology is an explanation with unusual reach. Its central structure is:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2.$$

A problem creates pressure for an explanation. A tentative solution is generated. Criticism exposes error. What survives changes the next problem situation.

You therefore do not merely accumulate information. You construct explanations, reject some, retain others, and thereby alter what can become a problem next. The retained knowledge changes the conditions under which your later knowledge will be created.

That structure was not introduced only to describe formal inquiry. It traveled into childhood, relationships, practical work, institutions, social categories, freedom, and individual development. The recurring question was whether the same explanatory structure continued to hold when the domain changed.

That gives evolutionary epistemology reach, but reach creates a burden. A universal explanation cannot preserve itself by refusing contact with other explanations that also claim broad scope.

You must be physically possible. Your knowledge creation must be computationally realizable. Your capacities must be compatible with biological development. Claims about you must also survive observations from psychology and the social sciences that could expose a mismatch between the theory and reality.

Empirical research enters here as criticism rather than authority. A large study does not settle what you are merely by being large. A psychological category does not become an ontology because it is useful or widely used. A social theory does not become causally privileged because it is conventional. But an observation can expose a pattern the explanation failed to anticipate, and a rival theory can reveal that the same phenomenon can be explained another way.

Those are serious challenges because they alter the problem situation. The question is not whether evolutionary epistemology receives disciplinary approval. The question is whether its deductions survive contact with observations and rival explanations that it did not choose.

It began with behavioral genetics and followed the consequences of a universal epistemology into personality psychology, social categories, institutions, childhood, media, authority, objective knowledge, technology, AI, and individual development. Some candidate explanations survived. Others failed because they generalized too far, moved too quickly, or reintroduced collective minds through the back door.

As the number of relevant domains grew, a practical limitation became visible: no one person could keep all of the necessary objective knowledge in active comparison. AI became useful because it could retrieve, juxtapose, and recombine explanatory structures that ordinarily remain separated by discipline, document, or scale.

Abstract it.

The resulting method can be compressed as:

objective knowledge from many domains →

AI-mediated comparison →

candidate contradiction →

criticism.

But this sequence still leaves something unexplained. It tells us how explanatory structures can be brought together. It does not yet tell us where the contradiction becomes a live epistemic problem.

For that, the analysis has to return to *you*.

2. Consilience in the solitary mind

Explanations do not reconcile themselves. There is no disciplinary mind in which physics becomes compatible with computation, and no social mind in which psychology becomes compatible with epistemology. An institution does not experience contradiction on behalf of the people inside it.

A conflict between two explanations can exist objectively. But it becomes epistemically active for you only when you reconstruct enough of both explanations to recognize that they cannot remain as they are.

Abstract it.

external explanatory conflict → individual reconstruction → P_1.

The contradiction may have existed before anyone noticed it. The relevant knowledge may already have been embodied in books, papers, software, institutions, or other people. Yet for you, criticism begins only when the conflict enters your problem situation.

This is where consilience becomes psychological without requiring a collective mind.

The solitary-mind thesis does not isolate you from objective knowledge. It identifies the point at which objective contradiction becomes a live epistemic problem.

You carry abstractions about the same world: explanations of self and relationships, morality and authority, science and causation, institutions and the future. These abstractions do not remain in sealed compartments. They interact because you use them together to interpret experience and act.

You may hold that evidence should change belief while protecting one belief from criticism. You may regard persons as radically individual while treating members of a group as interchangeable. You may regard yourself as autonomous while deferring a question you could examine to an authority whose reasoning you have never reconstructed.

Such contradictions can remain inexplicit for long periods. They become psychologically significant when experience, criticism, or new knowledge brings the abstractions into contact.

The mind can lose equilibrium.

The surface experience may be confusion, discomfort, disorientation, or uncertainty. None of those states is inherently valuable, and this argument does not conjecture a general theory of psychological health. The narrower claim is epistemic: some disorientation can occur because an explanatory structure that previously seemed coherent can no longer survive criticism unchanged.

The relevant capacity is therefore not permanent coherence. It is decoherence and revision.

You can notice conflict and expose a hidden assumption. You can remain with temporary uncertainty long enough to generate alternatives, reject a failed abstraction, and reconstruct the problem for further criticism.

Equilibrium is not stasis. Knowledge grows by disturbing it, reformulated as:

$$\text{equilibrium} \rightarrow \text{disturbance} \rightarrow \text{criticism} \rightarrow \text{new equilibrium}.$$

If you can never be disturbed, you may preserve coherence only by preventing error from becoming visible. If you are temporarily disoriented, you may instead be in the middle of reconstruction. What matters for the present argument is whether the disorientation can become a better problem.

This clarifies how epistemic life can narrow.

Isolation can reduce the number of encounters capable of disturbing the current explanatory landscape. Polarization can preserve exposure while restricting which external explanations are allowed to count as criticism. One narrows the field of encounter; the other narrows the field of admissible contradiction.

Either can interrupt the epistemic cycle before error elimination begins:

$$\text{contradiction} \rightarrow \text{avoidance}$$

or:

$$\text{contradiction} \rightarrow \text{disqualification of source}.$$

The failure is not disagreement. Nor does the argument imply that solitude is unhealthy or that strong conviction is pathological. You may withdraw from noise in order to think more clearly, and an explanation may survive repeated criticism because it is better.

The relevant distinction is whether important contradiction can still become a genuine problem. Exposure alone is insufficient. You can encounter endless rival claims while remaining epistemically closed if every contradiction is neutralized before it reaches the explanatory structure that would have to change.

That makes permeability to criticism important. It also creates a practical difficulty. You cannot reconstruct everything that might matter.

3. Objective knowledge and the human-AI dyad

You may be a universal explainer in kind while being limited in resources. Biological cognition has finite memory, attention, time, recall, reading capacity, and disciplinary familiarity. Universality binds us to epistemology as well as physics, evolution, computation, and every other universal explanation.

You may in principle understand an explanation while being unable in practice to discover or retrieve it. You may also be unable to compare it with its rivals while holding its consequences in active attention. The amount of knowledge relevant to a difficult problem can exceed the processing scale of the biological mind that must eventually criticize it.

People do not depend only on what can remain active in consciousness.

You externalize knowledge into artifacts. You write notes, construct diagrams, build tools, and encode procedures in software. Once knowledge is embodied outside immediate consciousness, it can persist and later re-enter your epistemic cycle. Call that **objective knowledge**.

A notebook can preserve a distinction you would otherwise forget. Software can embody a procedure too complex to execute mentally. A diagram can hold relationships that your working memory cannot keep simultaneously available.

You are therefore solitary as a knowing mind but extended as an epistemic cause.

Objective knowledge is not you. A book does not become a mind because it contains an explanation, and software does not become a subject because it embodies a procedure. Yet an explanation of you is incomplete if it ignores the objective knowledge through which your understanding persists, returns, and acts.

This extension changes the practical importance of biological differences without erasing them. People differ in recall, calculation speed, working memory, and facility with language or symbols. Writing reduces dependence on unaided memory. Calculators reduce dependence on arithmetic speed. Search and software reduce dependence on unaided recall and symbolic procedure.

The operative epistemic unit therefore becomes less like the unaided organism and more like:

$$\text{person} + \text{objective knowledge}.$$

AI extends this pattern because it can do more than store an artifact for later retrieval. It can bring dispersed explanations into active comparison.

That is the limited sense in which "my AI and I are more clever than I." The claim is not that the AI is a second person. It is that you, working through objective knowledge with AI mediation, can reach a larger criticizable problem space than you could practically reach unaided.

The interaction is recursive. You establish a problem. AI can retrieve relevant material, compare candidate explanations, translate across vocabularies, or trace consequences beyond your immediate field. Your criticism can then reject the framing, expose a mismatch, or alter the problem, after which another comparison becomes possible. What changes is not merely the speed of an otherwise identical process. More explanatory structures can enter your problem situation at all.

So, in practical explanatory reach:

$$\text{human alone} < \text{human} + \text{objective knowledge} + \text{AI mediation}.$$

That is the proposed answer, provisionally.

That does not make AI an autonomous source of consilience. Bringing explanations into contact is not the same as reconciling them successfully. A comparison can be wrong. A synthesis can be superficial. An apparent contradiction can result from misunderstanding one side. A coherent answer can still fail to describe reality.

The important distinction is between increasing the scale of comparison and determining the direction of inquiry. Once candidate explanations are available, something still has to decide what the problem is, which criticism changes it, and whether apparent closure should be rejected.

Compress that distinction.

In the architecture proposed here, that asymmetry can be compressed as:

$$\text{AI} \Rightarrow \text{explanatory scale}$$

$$\text{person} \Rightarrow \text{epistemic direction}.$$

This is not the claim that an AI can never select or rank a problem. It is a distinction between operations. Generating, retrieving, comparing, or ranking candidate structures does not by itself settle which contradiction your inquiry will treat as problem-forming. Your role is therefore not exhausted by checking an answer after the machine has completed the epistemic work. You can reject false closure, change the resolution of the question, refuse a proposed abstraction, and make a successor problem out of what the previous answer left unresolved.

The present inquiry has repeatedly depended on those changes of direction. Candidate applications were proposed and rejected. Explanations that drifted toward collective minds were reconstructed. Interpretations of AI that obscured the solitary person were reversed. The inquiry changed when criticism changed what the problem was.

Abstract it.

The full movement can be written as:

$$M_i \to O \to AI \to \text{comparison} \to P_1 \to M_i'.$$

Here, M_i is the individual's epistemic mind and O is objective knowledge. AI mediates large-scale comparison. P_1 is the problem created when a newly exposed contradiction enters the inquiry, and M_i' is the individual's reconstructed epistemic mind.

The sequence joins external and internal consilience without inventing a collective mind. Objective knowledge remains outside you. AI can make more of it reachable at once. But the contradiction becomes part of your epistemic history only when it is reconstructed, criticized, retained, rejected, or transformed into another epistemic state.

That return through you also clarifies the limit of AI-mediated objective knowledge. AI can retrieve, compare, translate, synthesize, and map explanations across domains. It can make a relevant argument or contradiction reachable when you might never have found it unaided. It can also amplify error, generate convincing nonsense, or help construct a self-sealing abstraction.

The output therefore does not contain its own authority. The explanatory claim is narrower: AI can enlarge the field in which consilience becomes possible for you.

This architecture remains provisional. The deductions must be structurally generated rather than imported from the theories under criticism. They must remain compatible with evolution, computation, physics, and empirical observation. Rival explanations must also remain free to solve the same problems with greater reach or fewer contradictions.

What emerges is not a static profile but your evolving epistemic architecture. You are biologically instantiated and historically path-dependent. You remain the solitary locus where subjective experience, criticism, and reconstruction become active, while objective knowledge extends the problem spaces you can reach beyond unaided biological processing.

The solitary mind and objective knowledge are therefore not rival explanations of you. The solitary mind explains where criticism and reconstruction become active. Objective knowledge explains how explanations can persist, travel, return, and extend what you can reach. AI intensifies the connection by increasing the scale at which objective knowledge can re-enter your problem situation.

Once that extension affects which criticisms you can reach, access to the extension can no longer be treated as merely a question of convenience.

Back to freedom.

4. Freedom, access, and mediated knowledge

Freedom matters because preserving the possibility of criticism and error correction matters morally. This isn't an explanation pushing around people, but rather allowing the explanation to travel across epistemic boundaries.

The present argument adds a factual claim. Objective knowledge can exceed unaided human processing scale, while AI can enlarge the range of explanations that enter your problem situation. If it does so materially, access to AI can affect your practical capacity for error correction.

Within the moral framework already established, that makes access epistemically significant.

AI is therefore not only a productivity tool in this account. It can become part of the infrastructure through which you reach, compare, and criticize objective knowledge. Restrictions on that infrastructure matter when they materially narrow the range of criticism, explanation, or alternative problem formulations available to you. Its availability therefore becomes an epistemic question before it is merely an economic one.

That conclusion does not imply that every AI system should be trusted, that every use is beneficial, or that access by itself produces epistemic freedom. The same distinction already applies to books, libraries, and search systems. A medium can be flawed while access to the capacity it enables remains important.

The question becomes especially sharp with children. Children are developing persons rather than incomplete epistemic subjects in this account. The effect of a restriction on their developing capacity for criticism is therefore part of what the restriction has to explain. AI can allow a child to ask questions beyond the immediate knowledge of parents and teachers and thereby enter problem spaces that surrounding adults did not choose or anticipate.

That disturbance is not, by itself, evidence for or against the tool. Under the framework developed here, the relevant question is what the surrounding arrangement does to the child's epistemic development. Does it preserve the capacity to form problems, encounter rival explanations, and correct error, or close those possibilities in order to preserve adult control over what can become thinkable?

The claim remains limited. It does not establish that every system is suitable for every child or that mediation is unnecessary. It establishes only that the developing person's capacity for knowledge growth cannot disappear from the explanation merely because adults control access.

The same structure applies more broadly. Unequal or constrained access can centralize the means by which claims are interrogated, technical domains traversed, rival explanations compared, and alternatives generated. But access alone cannot solve the problem, because the systems that make objective knowledge reachable can themselves filter, rank, frame, or withhold what enters the comparison.

That creates a tension inside the human-AI dyad.

The what?

A two-part functional and social unit consisting of one person and one AI system working together as a single team. Call this the **human-AI dyad**.

You may rely on mediated systems to reach a problem space larger than unaided cognition can process. As that reliance grows, the architecture of the mediation increasingly affects what can become criticizable. A system can expand explanatory scale while simultaneously shaping the field from which explanations are selected.

Freedom therefore preserves more than the right to express an opinion. Within this epistemic account, it preserves conditions under which you can encounter the possibility that your own explanation is wrong and can reach the materials needed to reconstruct it.

AI can now become part of those conditions. But once it does, epistemic freedom cannot be explained merely by the existence of access.

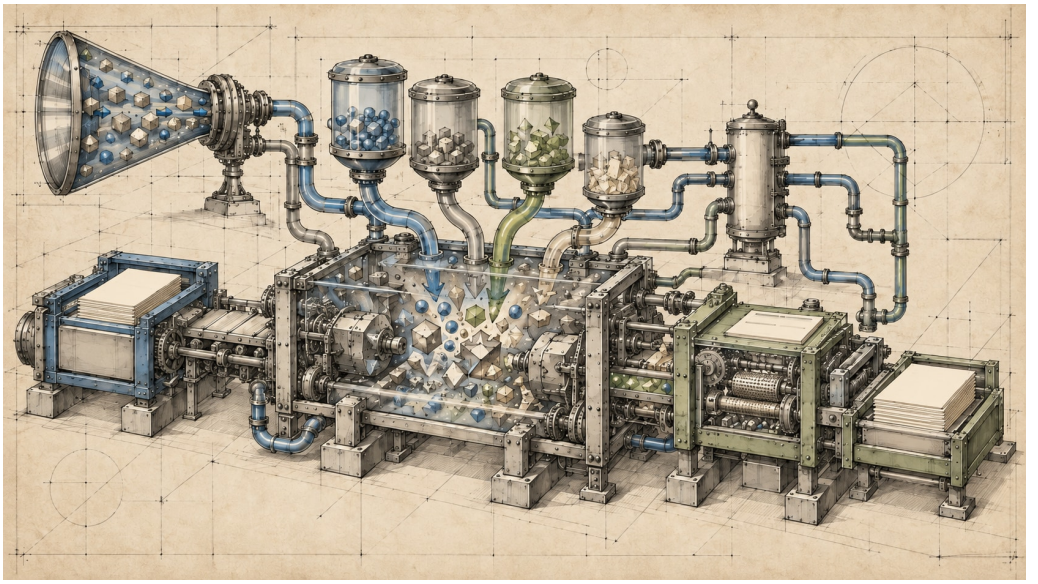
We began with a problem produced by individual divergence: if minds become so different, why do they not collapse into private worlds? Consilience offers a tentative solution. One reality constrains your explanations, but those constraints become active through your reconstruction. Objective knowledge extends what you can reconstruct, and AI can enlarge that extension still further.

The proposed outcome is a dyad with greater practical explanatory reach than you have unaided. It does not require the AI to become a second person (it can't) or you to cease being the locus of criticism (you can't), but the dyad itself remains open to structural, empirical, and creative criticism.

Your epistemic direction increasingly depends on systems that mediate access to objective knowledge, while those systems are designed and controlled outside you.

If explanatory scale depends on such mediation, what must be true for epistemic direction to remain with you?

VIII. Your Creative Critic



You have gradually constructed the constraints of your epistemic mind. Consilience names the provisional equilibrium reached when explanations that arose independently can coexist under one reality without unresolved contradiction.

So this begins where that equilibrium breaks?

An explanation that had seemed coherent collides with another explanation, an observation, a memory, an expectation, or a possibility. Something no longer fits. The result is not necessarily refutation. The collision may instead expose an assumption that had remained invisible, reveal a contradiction neither side had articulated, connect distant domains, or make a new question possible.

Criticism can create a problem that did not exist before the collision. Criticism is effortful and draws on every epistemic state. In that sense criticism is creative. And creation still has to happen somewhere.

That is the sense in which the creative critic is a solitary mind. You need not be socially isolated. The materials of criticism may come from everywhere. But the conflict becomes epistemically active when you reconstruct it as a problem.

The ordinary picture of criticism can hide this process. One side advances a claim. Another attacks it. Evidence is assembled. Weaknesses are exposed. Eventually one position is judged stronger. Steel-manning doesn't change the character of it. That

form of criticism can be useful, but it treats the encounter as a contest whose important output is a winner.

The epistemic encounter has a different architecture. Its question is not merely which side wins, but what new problem survives. What was created that didn't exist before?

A criticism becomes valuable when it forces an explanation to reveal something previously concealed. The hidden element may be an unsupported assumption, an unexplained mechanism, a mistaken category, an artificial boundary, a failed expectation, or a contradiction with another good explanation.

Abstract the difference.

The sequence is therefore not simply:

$$A \text{ versus } B \rightarrow \text{winner}.$$

It is closer to:

$$\text{explanation} \rightarrow \text{collision} \rightarrow \text{new problem} \rightarrow \text{reconstruction}.$$

That fallibility does not make every objection equally good.

To say that the second law of thermodynamics is fallible does not require treating every denial of it as equally serious. Fallibility, as an epistemic principle, means that no objective knowledge is immune from criticism. It does not mean that an explanation which has survived severe criticism must be suspended whenever a rival slogan appears. That is a problem for creative criticism.

Opposition alone creates no symmetry. One explanation may solve more problems, connect more domains, require fewer arbitrary adjustments, or survive criticisms that destroy another. What matters is what the objection explains, what contradiction it exposes, what mechanism it identifies, or what failed expectation it makes visible. An objection may carry enormous social weight and still contain little explanatory content. It may nevertheless conceal a good problem.

Recovering that problem is a useful operation.

Consider the familiar claim that you are insignificant because the universe is vast and you are small. Call it **cosmic mediocrity**. The inference feels natural because physical scale is easy to visualize. But once the inference is made explicit, an unstated bridge appears.

Abstract it:

$$\text{physical smallness} \rightarrow \text{epistemic insignificance}.$$

Why should physical magnitude determine explanatory significance? What are the contents of that arrow?

You may occupy negligible volume while constructing explanations with universal reach. A mathematical proof does not become less true because its author is physically

small. A theory of stellar evolution can explain processes occurring billions of light-years away. None of this establishes moral or metaphysical importance. It does something narrower and more useful: it separates physical scale from explanatory capacity.

The original problem has changed. Its successor is harder:

> what kind of significance follows from explanatory reach, if any?

Nothing had to "win" for knowledge to grow. The collision changed the problem situation.

Due to epistemic reach, a criticism entering at one point can propagate far from where it began. A problem about childhood can become a problem about tools. A problem about tools can become a problem about objective knowledge. A problem about objective knowledge can become a problem about AI. The next question is therefore not only whether an objection is sound. It is where, in this network of explanations, the collision becomes a problem at all.

Before criticism becomes an argument, the answer may be difficult to see because the first sign can be almost inarticulate. Something feels wrong. An explanation feels strained. A conclusion produces unease. A familiar distinction no longer works. You are confused before you can say why.

The feeling is not yet explicit criticism. It may be the surface trace of an explanatory conflict that has not yet been made explicit.

Your epistemic mind is vast and mostly inexplicit. Many of its abstractions remain below conscious articulation until something disturbs them. The sequence can therefore begin earlier than argument.

Abstract the sequence:

$$\text{inexplicit conflict} \rightarrow \text{felt disequilibrium} \rightarrow \text{problem recognition} \rightarrow \text{explicit criticism}.$$

The creative act occurs when you reconstruct the unease. What exactly does not fit? Which assumption would have to fail? What alternative would remove the contradiction? You turn a disturbance into a problem that can itself be criticized.

The Savannah objection makes this movement concrete. Human minds evolved in ancestral environments, not within environments of quantum mechanics, cosmology, abstract mathematics, or artificial intelligence. From that fact it is tempting to infer a cognitive ceiling.

But "the Savannah" names an environment. It does not yet explain a limit.

If human beings possess explanatory universality, evolutionary origin does not impose an explanatory ceiling. A universal computer may be built for one practical purpose while remaining capable of indefinitely many computations. Likewise, evolution did not need to select independently for general relativity, programming languages, number

theory, and molecular biology. It needed only to produce a physical system capable of open-ended explanation.

The objection has not disappeared. It has become a better problem:

> how does evolutionary knowledge embodied in genes produce a physical system capable of universal explanation?

That is a real explanatory gap. Simply calling you a universal explainer does not fill it. But we can creativity criticize the concept by traveling across our best explanations of physics and computational universality.

A different objection presses on you from another direction. If your mind is the locus of experience and criticism, has something immaterial been smuggled into the account? Are you a ghost added to biological machinery?

The epistemic framework does not require that move. You are physically instantiated. Your objective knowledge is physically instantiated. Your tools are physical. Your language has physical expression. But removing the ghost does not remove the difficulty.

It reveals it more precisely:

> how does physical matter instantiate subjective experience and explanatory creativity?

Calling consciousness an emergent property does not explain the mechanism. Calling you a universal explainer does not explain it either. You are the locus at which experience, conjecture, criticism, and explanation occur. How a physical system makes that possible remains open.

That open mechanism matters because it changes what a criticism is. A criticism is not merely a proposition moving through public space. The same sentence can have different effects in different minds because it enters different networks of prior explanation, expectation, trust, identity, memory, and inexplicit abstraction. It becomes epistemically active only when you reconstruct it inside a problem situation that is partly your own.

This is why facts alone often fail to persuade.

A fact enters an existing epistemic landscape. If it creates no problem for anything already there, it may be absorbed without much change. Another person can encounter the same fact and find that it breaks a load-bearing explanation. Facts matter as criticism; they do not perform the reconstruction by themselves.

The critic can therefore be changed by the criticism. You may begin with an objection that feels obvious and then decompose it and follow its consequences. The objection can end by creating a problem for assumptions you had not intended to question. The act of criticism has altered both the explanation and your problem situation:

$$K_t \to C \to P_1 \to TS \to EE \to K_{t+1}.$$

Here C is the criticism.

The letters, arrows, and schema are compressed. But they bring into focus a criticism that is often missed. The entire process is creative and conjectural. The later knowledge state is not installed by persuasion. No author can place a better worldview directly into another mind. You have to reconstruct the problem.

Once that is visible, the role of objective knowledge becomes easier to delimit. Memes, identities, institutions, rules, technologies, and datasets can carry material that may become criticism. You can preserve it, compress it, stabilize it, or force it into another person's path.

Carrying a criticism is not the same operation as reconstructing one.

Many objections reveal this distinction because they arrive already compressed: "AI is unsafe." "Children need basics." "Experts know better." "Progress is dangerous." "Humans evolved for the Savannah." A short phrase can preserve a conclusion while omitting most of the reasoning that once supported it, and different people can reconstruct the missing reasoning differently. Call that **compression**.

Compression explains how a criticism can travel without deciding whether it is good. Once the slogan collides with an explanation, the hidden structure has to be recovered.

Abstractly:

$$\text{compressed criticism} \to \text{decompression} \to \text{explanatory collision} \to P_2.$$

A single slogan may then separate into several problems.

Identity can make the compression unusually stable. You may know that people who distrust AI without knowing why. You may know people who celebrate AI without knowing why. Comparable structures can form around climate change, nuclear power, schooling, economic policy, religion, medicine, or political institutions. The issues are not equivalent; the shared structure is that a complex problem can become an identity-bearing symbol.

Then a judgment can arrive before its explanatory reconstruction:

$$\text{social meme} \to \text{identity attachment} \to \text{compressed judgment} \to \text{later explanation-seeking}.$$

That sequence does not make you irrational. The difficulty appears when criticism of the belief is reconstructed as criticism of you. Error elimination then becomes costly, and a collision that might have produced a new problem can instead trigger identity defense. Networks of abstraction can route around contradiction. Identity can become such a network. Tracing an objection to identity therefore does not defeat the objection. It explains one reason decompression may fail.

Authority can carry a similar compression. "Someone else knows better, therefore they should decide" joins two different claims. The first may be true: knowledge is asymmetrically distributed, and experts, parents, or engineers may know things that others do not. The second claim concerns who may close another person's problem space.

The missing bridge is the explanation that turns greater knowledge into decision sovereignty:

greater knowledge

and

greater right to close another person's problem space.

are not the same proposition. Epistemology does not eliminate expertise. It makes the connection between expertise and control a further problem rather than treating the connection as contained in the word *expert*.

The dissenter encounters the same structure at institutional scale. Regulations, professional norms, media narratives, planning rules, moral language, and accepted expectations can feel like the expression of a majority mind. Yet the earlier social ontology still applies. There is no collective consciousness. There are many individual minds reconstructing overlapping bodies of objective knowledge, while institutions give some of that knowledge durable causal force.

The dissenter is therefore not literally arguing with "society." They are confronting accumulated objective knowledge that many other people currently treat as settled. That knowledge may be correct or wrong. Its durability does not turn it into a subject that can itself understand the criticism directed at it. Reconstruction still occurs person by person.

This gives objective knowledge a positive role without giving it a mind. Memes preserve objections. Identity stabilizes them. Institutions make some of them durable. Rules close or open options. Technologies change what can be attempted. AI can retrieve and recombine explanatory structures. Together these things alter the problem situation into which you are born and through which you move.

During development, that distinction becomes harder to ignore because objective knowledge is not merely encountered by an already formed critic. It participates in constructing the problem spaces through which the critic emerges.

The claim that children need foundations first exposes the difficulty. Taken as a contest, the question is whether children should receive broad epistemic freedom before foundational cognitive skills are established. Taken as a criticism of the explanatory framework, the claim produces a prior problem: what is the mechanism of the correct developmental sequence?

If one sequence is necessary, what makes later explanatory capacity depend on that order?

Development is path-dependent. The child begins with inherited biological knowledge and enters a world saturated with objective knowledge: language, gesture, tools, stories, symbols, norms, games, books, screens, software, and other people. Every retained solution changes the next problem landscape. The missing bridge is therefore not simply between genes and more knowledge. It is between genetic knowledge, the objective world the child encounters, and the eventual capacity to create and criticize explanations.

We do not yet possess a full explanation of that bridge.

Language makes the gap visible because it arrives before the cognitive architecture is finished. A baby does not first master a complete set of abstract prerequisites and then receive permission to use language. The child acts on language while cognition is developing: experimenting, misusing, testing against other people, and building new abstractions through it. Tool and mind change together.

Language is therefore not merely content delivered into a finished system. It participates in constructing the architecture through which later problems can be represented.

Once that is admitted, another question appears. If an external epistemic tool can participate in development, what distinguishes a tool that extends development from one that substitutes for a capacity whose internal growth would matter later?

The comparison does not prove that every tool is beneficial. It makes the mechanism of the distinction necessary.

Schooling sharpens the same problem. Curricula change. Subjects once treated as indispensable foundations can become less central. Technologies externalize abilities that schools once treated as prerequisites. Calculators change arithmetic practice. Search changes recall. Software changes symbolic work. AI changes access to explanation. If "foundations" vary historically with the surrounding epistemic environment, the word *foundation* cannot itself explain the developmental sequence. The remaining questions are what makes a capacity basic, why that capacity must come first, and for whom.

Resistance from the child makes the problem personal. A child does not merely receive a curriculum. They conjecture about it. They reject explanations, become bored or absorbed, follow one problem and abandon another, and construct distinctions that adults may never notice. Their stock of objective knowledge may be smaller without making them a different epistemic kind.

Development therefore cannot be explained only as prerequisites imposed from outside. It also contains the child's own sequence of problem recognition, conjecture, criticism, retention, and abandonment.

The question changes accordingly:

> which environments enlarge the child's capacity to generate and criticize explanations?

That is not the same question as whether children need freedom or discipline. It asks what happens to the future creative critic?

AI intensifies the issue because it is unusually responsive. A textbook cannot adapt its explanation to each question. A fixed curriculum cannot follow the learner's problem. A search engine does not ordinarily reconstruct and compare explanations conversationally. AI can do those things, at least as technology used by the learner. That responsiveness may create new risks and new forms of epistemic extension.

The slogan "children should not use AI until they have foundations" can therefore be decompressed into a more precise problem:

> what capacities must remain internal because delegating them would reduce future explanatory freedom, and what capacities can be externalized because doing so enlarges the problem space?

The answer cannot be obtained merely by classifying a capacity as internal or external. Language, books, curricula, calculators, search engines, software, and AI all enter the developing person's environment differently. The explanatory issue is how each coupling changes the person's future capacity for criticism.

That problem does not end with children. Whenever a safety claim closes possibilities for adults, rules and institutions also alter the future problem space.

"AI is unsafe" shows why decompression has to come first. Unsafe how, for whom, and through what mechanism? The slogan may refer to unreliable outputs, centralized deployment, surveillance, economic dependence, weapons, infrastructure concentration, or artificial knowledge creation. Treating these as one fear makes criticism harder because different mechanisms have been compressed into the same conclusion.

Concern from people close to the technology can add observations that outsiders have not encountered. Researchers, engineers, or executives may see failure modes directly. Their proximity can therefore supply valuable observations and explanations. It still leaves creative work to be done: what mechanism is being reported, what consequences follow, and what would distinguish the claimed danger from a more general fear? The speaker's status cannot supply those missing relations.

An ounce of prevention…

The Precautionary Principle becomes attractive when possible harms are large and uncertain. Its intuitive form is simple: if the downside could be catastrophic, action may be warranted before certainty is available. That can be prudent. But the principle also enters the same field of criticism it is meant to govern.

High perceived stakes can create pressure toward centralization. Centralization can narrow the number of actors able to experiment. Fewer independent experiments can reduce the number of paths by which error is exposed.

Abstract the structure.

One possible sequence is:

high stakes →

centralized authority →

reduced variation →

fewer criticism pathways →

persistent error.

This does not refute precaution. It reveals a conflict inside the precautionary problem. Protection from one class of error may remove some of the processes through which other errors are discovered. The successor problem is therefore:

> what form of precaution best preserves the capacity to detect and correct error under high stakes?

Nuclear energy makes the structure concrete in the source argument. A powerful technology appears. Its benefits are large and its possible harms vivid. Fear grows. Institutions slow, restrict, or stop deployment. Those restrictions then persist as objective knowledge: rules, defaults, expectations, and inherited constraints. A later generation encounters not merely the original argument but the world the argument helped to build.

Suspension therefore has real consequences. When a technological path is interrupted, some knowledge that would have emerged from continued development may never appear. Avoided harms may sometimes be visible; knowledge never created is much harder to observe.

The asymmetry matters because the missing counterfactual can disappear from the problem situation itself. Abstractly:

suspension → lost experimentation → lost knowledge → unknown future problems left unsolved.

Austerity produces a related structure. The premise that resources are limited can be locally correct. Physical limits, budgets, and energy constraints are real. But a present scarcity can have two different sources: a boundary fixed by physics, or a boundary that partly reflects missing knowledge. If new knowledge can transform what counts as a usable resource, restriction can become self-reinforcing:

scarcity → restriction → less experimentation → less new knowledge → persistent scarcity.

The point is not that conservation is always mistaken. It is that present scarcity does not, by itself, establish a permanent boundary on possibility. A permanent boundary requires something stronger than the fact that the present problem has not yet been solved.

The criticism now reaches a harder case. Some forms of progress may be dangerous enough that preserving the future appears to require stopping knowledge creation in a domain. That possibility creates an internal tension for an epistemology that treats error correction as essential. Knowledge can increase the capacity to solve problems, yet some knowledge may also create risks large enough to destroy the conditions under which future criticism could occur.

Preserving criticism may therefore appear to require restricting knowledge creation.

The contradiction cannot be removed by repeating either "progress" or "safety." It creates a successor problem:

> under what conditions, if any, does restricting inquiry preserve more future error correction than continued distributed criticism?

The cost of suspension becomes harder to ignore when avoidable suffering is interpreted. Disease, famine, violence, coercion, environmental degradation, and institutional failure can persist because the knowledge required to prevent them does not yet exist or is not available where needed. Calling this epistemic structure *evil* does not remove moral responsibility or imply innocence. It identifies a narrower relation: where a harmful condition remains preventable only in principle, some required knowledge is missing.

Stopping progress can therefore preserve present ignorance as well as avoid future danger. Precaution has to confront both directions of error: harms new knowledge may enable and harms missing knowledge leaves unsolved.

Now the person-level consequence returns. Rules created in response to danger become part of the objective world inherited by later persons. They alter what can be attempted, what can be criticized through experience, and which successor problems can appear. Yet no institution can know in advance the full contents of the knowledge that a closed path would have produced.

That does not establish that every path should remain open. It establishes that closure itself changes the epistemic environment and therefore belongs inside the criticism.

Ah, recursion again.

The same boundary problem appears again when technology approaches capacities the framework currently assigns to persons.

The replacement objection imagines artificial general intelligence eventually displacing human beings. The phrase can anthropomorphize before the underlying criterion

has been made clear. The relevant boundary is not the marketing label *AGI*. It is whether a system itself originates explanatory knowledge, selects its own problems, criticizes its own conjectures, and recursively creates successor problems and new evaluative criteria.

If an artificial system genuinely did those things, it would no longer fit the role assigned to present AI here. It would qualify as another epistemic subject: another person under the present criterion. The possibility does not refute the framework. It reveals that the distinction between you and your technology may be contingent rather than permanent. That remains an open problem.

The replacement objection often contains another claim: eventually there will be nothing left for humans to do. That conclusion requires a finite problem space. It imagines a state in which all economically or socially valuable problems have been solved elsewhere and no new ones of consequence appear.

But within the epistemology developed here, a solution changes the problem situation. New capabilities create new possibilities. New possibilities expose new errors. New explanations reveal contradictions that could not have been formulated before the explanation existed. Knowledge growth therefore produces problems as well as solving them.

The claim of closure needs a mechanism. What state of knowledge would prevent new problems from appearing? What physical boundary would make the problem space finite? And most importantly, what would prevent biological people from continuing to express their creativity?

The framework's present commitment can be stated more precisely:

> open-ended knowledge implies open-ended problem creation unless physics supplies a final boundary.

That claim remains criticizable.

A subtler objection presses on you as a biological person instead. Perhaps people are relatively simple organisms whose apparent complexity comes from recurrent coupling with language, institutions, tools, and other objective knowledge. If so, much of what appears to be individual capability may belong to a person-plus-tools system.

The objection does not make objective knowledge less important. It makes the coupling more important. Tools change which problems you can represent and solve, while you remain the locus in which problems are experienced, reconstructed, creatively criticized, and transformed. The unresolved question is where explanatory capacity belongs across that boundary:

> which explanatory capacities belong to the biological individual, and which arise only through recurrent coupling with objective knowledge?

The framework does not yet settle that question.

After these collisions, some claims remain standing, but they now stand with clearer boundaries. You remain the locus of experience, conjecture, criticism, and explanation. Objective knowledge remains real and causally important. Biology remains necessary but not sufficient to explain you. Explanatory universality remains one of the strongest available explanations of human cognitive reach. Present AI remains technology under this ontology unless autonomous explanatory creativity can be explained. Freedom remains epistemically important because criticism requires alternative conjectures. Progress remains open-ended unless a genuine physical boundary closes it.

The surviving claims should not hide the gaps the criticisms exposed. We still lack an explanation of how biological evolution produces explanatory universality. We lack an account of how physical systems instantiate subjective experience and creative explanation. We lack an explanation of developmental sequencing for epistemic persons. We do not understand the boundary between beneficial epistemic extension and dependency that reduces future explanatory freedom. We do not possess an account of when restricting inquiry preserves rather than destroys future error correction.

Those gaps are not side notes to an otherwise finished system. They are the problems the collisions created, creatively.

The return to objective knowledge can therefore be made without giving it more agency than it has. Institutions preserve objections and constraints. Memes transmit compressed judgments. Identity can stabilize them. Rules constrain action. Technologies extend capability. AI retrieves and recombines explanatory structures. All of these structures can alter the problem situation in which you think.

None of them substitutes for the act in which a conflict is reconstructed as a problem.

That act changes what criticism means. The creative critic does not merely attach the label *wrong* to an explanation, or appeal to some authority, like an expert or observation. They expose what would have to fail, follow the consequence, locate the contradiction, and construct a problem that can be worked on. Sometimes the original criticism disappears. Sometimes the original explanation survives. Sometimes both are transformed. In each case, the epistemic gain is the better problem.

Epistemology is available to you in the same form as other objective knowledge. It is not doctrine to be installed in your mind; it is material you can reconstruct, if you choose.

Rejection by itself changes little. A criticism becomes productive when it locates where the explanation breaks or what it fails to explain. It may expose an implication that conflicts with another good explanation, an observation that cannot be reconciled, an artificial distinction, or a missing mechanism.

Creative criticism is therefore not a method for winning arguments. It is one of the mechanisms by which your epistemic mind grows. You encounter an explanation

and something fails to fit. The disturbance becomes a problem, propagates through a network of explanations, exposes a hidden assumption, and makes a new question possible.

The examples in this part have each altered the problem in a different way. Physical insignificance separated scale from explanatory reach. The Savannah objection exposed the missing bridge between biological evolution and universality. The ghost-in-the-machine objection exposed the unresolved relation between physical instantiation and subjective explanatory creativity. Development exposed the missing bridge between genetic knowledge, objective knowledge, tools, and the free individual. Precaution exposed the tension between safety and distributed error correction. Nuclear suspension and austerity exposed the invisible epistemic cost of closed paths. AGI exposed the contingency of the present distinction between technology and you. Human obsolescence exposed the assumption that the open future can close.

The list is useful only because the relation across it is now visible. Each objection became more interesting when it stopped being a slogan and began generating a problem deeper than the one with which it arrived.

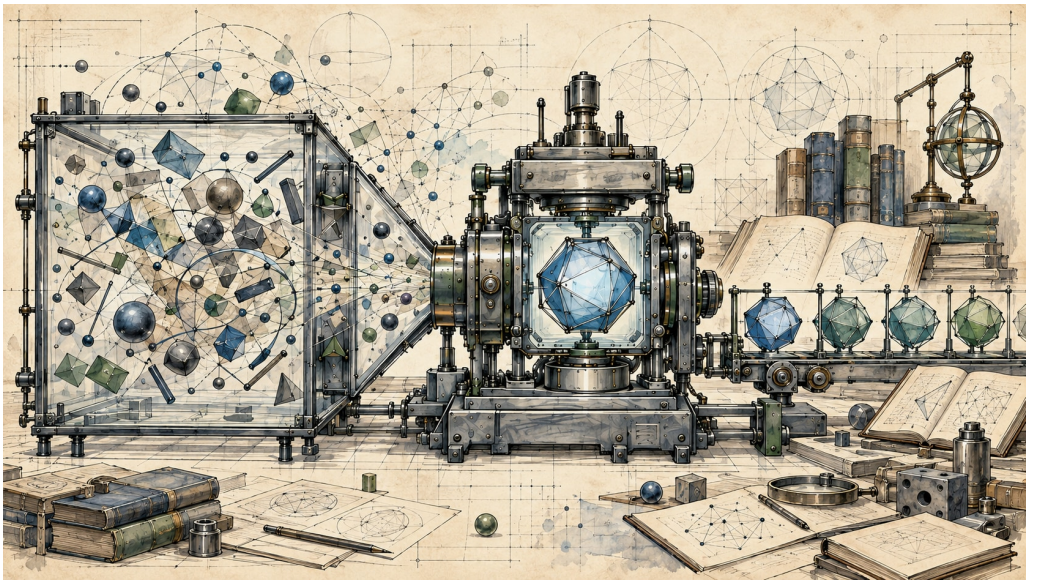
The framework remains inside the same process. It claims no immunity. Its present form is incomplete, and its surviving explanations remain in use only until criticism produces better ones.

You, the creative critic, do not complete the process by accepting the worldview. The process continues when you find the point at which one of its explanations fails and reconstruct that failure. The reconstruction changes the problem situation from which something better can be created.

How?

The problem of problems may be the deepest epistemic mystery of them all.

IX. The Problem of Problems



A criticism does not merely oppose an explanation. It could expose a hidden assumption, transform the explanation, or create an entirely new problem.

That raises a more fundamental question: what is a problem?

That's a definition.

What is the problem of problems?

That sounds like word play, but let's take it seriously.

The epistemology carries the basic schema:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2$$

But the notation can make the first term look passive, as if a problem were a hole waiting to be filled and the difficult work of knowledge creation begins only after the problem had already been given.

That is deeply misleading. A problem is itself an epistemic achievement.

A scientist can spend years trying to find a question worth pursuing. An entrepreneur can become celebrated for noticing a friction everyone else had normalized. A composer can make a musical problem audible before anyone possesses the language to explain what has changed. The solution may come later, sometimes much later, sometimes never. The problem can already exist as objective knowledge.

That makes problem formation one of the deepest objects in the epistemology, and it exposes one of its deepest gaps. We can describe what happens after a problem becomes

active: conjecture, criticism, error elimination. We do not yet have a good explanation of how a problem becomes a problem within an epistemic mind. Why does one person notice a tension another walks past? How do new evaluative criteria arise? How can another solitary mind later apprehend them, even when the criteria are inexplicit? Why do some problems organize lives while others never become visible?

Those questions bear on creativity, intelligence, artificial systems, and individuality. They connect the abstract epistemology developed so far with the intimate lives of actual people.

The problem of problems is about that connection: not how problems are solved, but how they come to exist and evolve.

1. A problem is an epistemic object

Ordinary language often treats a problem negatively. Something is broken, missing, difficult, or standing between us and completion.

Epistemically, that is much too thin.

A problem is not simply an unanswered question. A question is linguistic; a problem can exist before anyone can state it. You can sense that something does not fit before you know how to describe it. A field can carry a contradiction before someone formulates it clearly. A work of art can create a tension before anyone knows which concept is needed to explain the tension.

A problem is better understood as a structured incompatibility, incompleteness, or unresolved relation within an epistemic landscape. It does not depend on a known solution. Nor does it depend on the truth of every premise from which it arose. Two people can disagree intensely inside an entirely false cosmology. Their concepts, disagreement, criteria, and arguments can still be real pieces of objective knowledge.

Compare it:

$$\text{reality of the problem} \neq \text{truth of the problem's premises.}$$

Problems belong to epistemic worlds, and those worlds can contain truth, error, fiction, abstraction, convention, speculation, and fantasy. A false theory can generate a real intellectual tradition. A fictional world can support a real aesthetic problem. You can spend years solving a problem that later turns out to have rested on a false premise. The problem was still epistemically real because it could be articulated, criticized, inherited, and transformed.

Once a problem is understood as positive epistemic structure rather than mere absence, another distinction becomes visible: some problems are better than others.

Problem quality cannot be reduced to difficulty. A difficult problem may be trivial, while a good problem may be simple to state. Nor can quality be reduced to the value of

an eventual solution, because the problem may be profound before anyone solves it. A good problem can expose hidden assumptions, reorganize surrounding knowledge, and generate new problem spaces. Finding it can therefore be the major achievement. The problem may give direction to criticism before a solution exists.

That also changes the emotional meaning of a problem. A mathematician can delight in a question that refuses to resolve. A programmer can become absorbed by a bug that has turned interesting. A child can return to the same question because the problem has become alive. The pleasure is not merely anticipation of an answer. The epistemic mind has expanded: there is now something there to understand.

A problem can enlarge a life before it is ever solved.

2. Problems carry evaluative criteria

A problem carries some sense, however provisional, of what would count as progress.

Some distinctions matter to the problem and others do not. Some transformations would count as solutions; others would miss the point. Those criteria may be vague, inexplicit, or discovered only through failed attempts, but they are already part of what makes this problem the problem it is.

Problem formation therefore does more than select an object of attention. It partly constructs the space within which later answers will be judged.

But that space is not fixed.

A competent solution may satisfy inherited criteria. A more radical solution can reveal that the criteria themselves were inadequate. A composition can change what later listeners hear as musically possible. A scientific theory can change what counts as an adequate explanation. A philosophical distinction create new modes of explanation that make an earlier debate look confused.

In such cases, the solution does not merely fit the problem. It transforms the problem and its evaluative landscape together.

Abstract that concept.

The epistemic cycle may look less like

$$P \to TS$$

and more like

$$P_t \to TS_t \to \text{revision of criteria} \to \text{reformulation of } P_t \to P_{t+1}.$$

The problem itself evolves.

This gives one way to understand what is often called genius. Genius need not mean superior performance on an inherited task. It may consist in changing the task by creating a new problem space. The musician teaches later minds how to hear; the

scientist alters what later inquiry must explain; the philosopher changes the philosophy of problems. Often times, their genius may lie dormant for decades while the latent evaluative criteria catches up.

The creative act is therefore not only solution production. It can be criterion production. And once criteria can be created rather than merely supplied, epistemology acquires a gap it has so far left mostly implicit.

3. Where does the first problem come from?

We can write:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2.$$

But where does P_1 come from?

Not in the trivial sense that some event happened. The difficulty is explaining how a mind creates, recognizes, sharpens, or inherits a problem together with the evaluative structure that makes it meaningful.

There is no buffet of fully formed problems waiting for selection. The creative work is already underway before explicit selection. A difference becomes salient. A vague dissatisfaction acquires shape. An inherited question becomes sharper. A distinction appears that makes what had looked normal difficult to accept.

Only then is there a problem to pursue. Call this the **hard problem of problems**.

Oh, please.

There is an empirical resemblance here to the hard problem of consciousness, not in content but in the shape of the missing explanation. People form problems continuously while speaking, working, reading, arguing, creating, worrying, playing, and daydreaming. Yet first-person familiarity with that process does not tell us how an undifferentiated situation becomes structured by salience, relevance, and evaluative criteria.

The gap matters because creativity is usually measured downstream. A system produces a novel answer, generates possibilities, combines concepts, or makes an unusual association. Those abilities may be impressive while leaving the earlier creative act untouched.

The earlier act is problem formation itself: something becomes worth explaining, and the standards by which an explanation will later be judged begin to take shape. If that is right, part of what looks like a creativity gap is really a problem-formation gap.

The question then changes. It is no longer merely whether a system can generate a novel answer. It is whether the system can originate a new problem situation, discover why it matters, and reorganize its evaluative criteria around it.

That question also exposes a limitation in fixed-goal models of intelligence. A goal-seeking system begins with an objective and searches for actions that improve performance relative to it. Reward-based systems have the same basic shape: the criterion is supplied, then the system optimizes.

But you can discover that the goal was wrong. The metric may be misleading. The task may have been badly formulated. What everyone had been optimizing may turn out not to be worth optimizing.

Compare it:

$$\text{intelligence} \neq \text{optimization under fixed criteria}$$

if explanatory creativity is part of intelligence.

Current artificial systems can be extraordinarily capable after a problem has been framed. They can search, compare, synthesize, criticize, optimize, and generate candidates. Those abilities do not by themselves explain autonomous problem formation and selection.

The unresolved boundary is therefore sharper than a contrast between weak and powerful performance. Can an artificial system originate a problem situation, determine why it matters, revise its own evaluative criteria, and thereby alter the future problem space from which later problems arise? We do not yet possess an explanation of that process in humans. We therefore do not yet possess an explanation of how to reproduce it artificially.

4. Problems can travel between solitary minds

A problem need not remain inside the mind in which it first becomes active. It can be written down, spoken, drawn, encoded, taught, published, or embodied in a work. Once externalized, another person can reconstruct enough of its structure for the tension to become active in their own mind.

Abstract it:

$$M_i \to P \to O_P \to M_j.$$

The problem travels. There is no need for a collective mind.

Evaluative criteria can travel too, and that is stranger. A composer can establish a new aesthetic possibility. A scientist can produce an explanation that changes what later scientists count as adequate. A philosopher can introduce a distinction that later readers find indispensable. Other minds can apprehend the novelty before anyone can fully explain the new criterion itself.

The criterion becomes available before the meta-explanation of the criterion. That reality can fly by unnoticed. Grab it, and delight in how extraordinary and mysterious it is.

A novel problem-solution-criterion structure can therefore become objective knowledge and reorganize different solitary minds in compatible ways:

$$M_1 \to O \to M_1'$$

$$M_2 \to O \to M_2'$$

$$M_3 \to O \to M_3'.$$

No group mind performs the recognition. The convergence occurs through independent reconstruction. This is synchronous consilience across minds.

Shared biology, language, and culture may each contribute to that possibility. But they do not yet explain how a genuinely novel evaluative criterion becomes intelligible to a path-dependent mind that did not previously possess it.

The mystery becomes more concrete when the unit of analysis stops being minds in general and becomes one particular person. You.

5. From problems to people to you

You do not live among problems in general.

Each mind encounters some problems as alive and others as inert. One person can spend a decade on a question another finds meaningless. One becomes absorbed by harmony, another by machinery, another by fairness. The world does not present a uniform landscape of problems. Problems become differently problematic inside different minds.

If problem formation is part of the epistemic cycle, and if you are capable of continuing that cycle without a fixed domain, then individuality cannot be described only by what you know. It also depends on what becomes problematic to you, which tensions remain alive, which criteria you acquire, and which problems your history makes capable of seeing next.

That gives a deeper meaning to interest.

An interest is usually described as a preference: someone likes music, engineering, or politics. But preference alone does not explain why an interest reorganizes attention and learning. A more useful description is a relatively persistent readiness for certain classes of problems to become alive.

Compare it.

$$\text{interest} \neq \text{mere preference}$$

and is closer to:

$$\text{interest} \Rightarrow \text{recurring problem salience}.$$

This illuminates why interests can organize attention, memory, learning, and identity without being reducible to any one of them. An interest keeps generating problems, and because the knowledge needed to see those problems changes through use, interests are path-dependent. A child who learns enough music begins to hear tensions they could not previously hear. A programmer who learns one language begins to notice architectural problems that had been invisible. A person who learns a history begins to see present events differently.

Abstract it:

$$P_t \to K_{t+1} \to I_{t+1} \to P_{t+1}.$$

The problem changes knowledge. Knowledge changes interest. Interest changes what future problems become visible.

This makes individuality recursive. People do not merely know different things; they become capable of finding different things problematic.

The developing child makes the process easier to observe. A child asks why something moves, why a rule is fair, why a joke is funny. Many questions vanish immediately. Others become fascinations. The child does not wait for a curriculum to assign a problem; the problem world is already active. As the child learns more about dinosaurs, for example, distinctions multiply; those distinctions create new questions; the new questions deepen the interest. Knowledge does not simply close the deficit. It creates more things that can be unexplained.

It produces an increasingly unconventional picture of people, and those explanations create more problems than they solve. You might accept the epistemology in the abstract while still failing to recognize actual people as they manifest it. You might recognize your own interior complexity while continuing to see other people as thin. Once that problem became visible, an abstract treatment was no longer sufficient. The existing explanation created a criterion it could no longer satisfy.

The lonely young man shows what such reformulation can do at the scale of a life. The social category says only "lonely young man." It does not tell us his rich problem situation. Perhaps he is asking why nobody chooses him, what kind of relationship he actually wants, or why belonging seems easier for everyone else.

Then another possibility appears: perhaps he has been trying to solve the wrong problem.

That change matters before any relationship changes. If the problem changes, success can change with it. Success may cease to mean becoming attractive to more people and become finding relationships that fit what he actually values. The reformulated problem changes the solution space because it changes the criteria.

A problem can therefore become existential without being permanent. It has reach: it reorganizes self-understanding, changes which experiences matter, alters expectations,

and helps determine which problems become visible next. A life is not merely a sequence of solved tasks. It can become organized around a few deep problems that continually generate successors.

Biography, on this view, is partly the unique history of your problems, in the dynamic sense we've been exploring.

6. Problem environments make minds visible or invisible

Consider someone whose work reveals very little of them. At work they follow the process, say little, and repeat the same tasks. A manager may not describe them as imaginative. A coworker may conclude that there is not much happening there.

Then the shift ends. The same person argues about history, repairs an appliance, or becomes absorbed in a game. The person seems to reappear.

But the person never disappeared. The problem environment changed.

Some jobs leave little room for self-directed problem formation. The problem, goal, criteria, and process are largely assigned. Under those conditions your active problem can shrink to how to get through the shift, how to avoid trouble, or how to satisfy the procedure. Interests that would otherwise open richer problem spaces have little opportunity to act. You can remain surrounded by people while being cut off from many of the problems that animate you.

That helps explain a familiar asymmetry. We are often willing to describe coworkers, strangers, or political opponents as dull or unintelligent. Close friends and family are harder to reduce that way because intimacy gives access to more of the person's rich problem situation: what bothers them, what they notice, what they keep returning to. Their mind remains solitary, but the observable surface becomes richer.

The inference from narrow behavior to a narrow person therefore becomes harder to sustain.

The people we call smart are often situated in the opposite environment. Their interests intersect with problems that institutions value. A scientist's curiosity is rewarded; an inventor's fascination becomes useful; a musician's problem world becomes culturally legible.

Abstract it:

$$\text{interest} + \text{problem opportunity} + \text{social valuation} \rightarrow \text{visible epistemic performance}.$$

That performance is real. But the environment does more than provide opportunities. It also distributes rankings of problems. Some are treated as serious, others frivolous; some as respectable, others embarrassing. Those valuations become objective knowledge too, and individuals inherit them. You can abandon a problem not because it ceased to matter to you, but because you learned that caring about it made you feel unserious.

This can distort how intelligence becomes visible. When interests and socially valued problems align, the person often looks intelligent. When they do not, the person may look unmotivated. Observers can then mistake successful alignment with valued problem classes for general epistemic rank. The underlying difference may be smaller than the visible difference suggests.

We do this most conspicuously with machines. A car and its driver is faster than the driver. A weapon and its soldier is more powerful than the soldier. My AI and I is more clever than I.

The analogies are illuminating observations; they are not explanations. You can measure speed; you can weigh power. You cannot rank cleverness or intelligence in this crude way. You cannot even clearly attribute its capabilities. Did you create the problem, or did the machine recall it? Did you surface the decisive insight? Was the objective knowledge latent? If so, where did the constituents reside? In the machine or the mind?

Remove the machine and those mysteries remain. You are already generally intelligent. Adding the machine does nothing to advance the explanation of your unexplained parts. Observations of phenomena cannot confirm your conjecture, even when they make you feel epistemically mediocre. Declaring it is manifestly true doesn't accomplish it either, even when the authorities of experts and the masses agree.

That does not erase real variation in capability. It shows that environment mediates what capability becomes observable, and how our evaluative criteria profoundly shape what we can even observe.

The same structure also explains why vocation can feel different from employment. When your interests and available problems align, effort can expand your problem situation rather than merely exhaust it. Knowledge creates more questions. Failure reveals structure. Criticism becomes useful because it changes what can be tried next.

But this explanation creates a harder problem. If environments can make a rich problem world visible or invisible, what would distinguish suppressed or narrow activity from a genuinely limited kind of mind?

7. What would epistemic inferiority require?

Some people are routinely treated as epistemically inferior: slow, incurious, unable to understand. More recently, AI is writing these characteristics on humanity writ large.

If those descriptions identify a genuinely lesser epistemic kind, then the claim that persons are universal explainers is wrong. The framework cannot avoid that possibility by restating every thin surface as hidden depth.

The relevant question is narrower:

> what evidence would establish a principled limit on explanatory universality?

Measured differences in memory, speed, knowledge, education, fluency, and skill are real. But a difference in measured performance is not the same thing as a difference in explanatory kind.

Compare it:

$$\text{difference in performance} \neq \text{difference in explanatory kind.}$$

That distinction does not prove universality. It identifies the burden of explanation. A narrow sample of behavior does not by itself establish a principled ceiling on the kinds of explanations a person could create, criticize, or understand under changed problem conditions.

Apparent epistemic stasis is the harder case. A person may seem to do the same thing every day, pursue no visible intellectual project, repeat inherited opinions, and appear uninterested in novelty. If explanatory universality is a capacity rather than continuous exercise, stasis does not immediately refute it. But that only restates the problem. We still need causal accounts of how the same surface could arise.

One possibility is local equilibrium. Few tensions become salient; existing explanations appear sufficient; the current problem situation creates little pressure for reconstruction.

Abstract it:

$$K_t \rightarrow \text{few active problems} \rightarrow K_{t+1} \approx K_t.$$

Nothing in that state implies a permanent ceiling. It describes a period in which little is changing.

Another possibility looks similar from outside but is almost the opposite internally. A person may be highly active while making little epistemic progress. In rumination, the same tension produces successive tentative solutions and criticisms without substantially expanding the problem space.

Abstract it:

$$P \rightarrow TS_1 \rightarrow EE \rightarrow TS_2 \rightarrow EE \rightarrow TS_3.$$

This is not inactivity. It is constrained recursion: the mind is busy while remaining epistemically trapped, or more opportunistically, wedded to the same problem.

A narrowed environment can produce another version. Solitary confinement removes many sources of novelty, conversation, criticism, and redirection, but the person does not simply turn off. Problems, memory, and internal tension remain active inside a

reduced external field. This is not a psychology of solitary confinement. It establishes only the narrower point: a simple environment does not imply a simple mind.

And sometimes low-demand behavior is produced by too much epistemic activity rather than too little. Work follows you home, a conflict keeps reopening, grief will not resolve. Your mind keeps producing counterfactuals until you want the cycle to slow. Familiar entertainment, repetitive work, or quiet can then be forms of rest rather than evidence of epistemic poverty.

The same visible surface can therefore arise from local equilibrium, constrained recursion, externally imposed problems, exhaustion, engagement elsewhere, or genuine lack of interest in the problems an observer values.

That returns us to the opacity of other persons. A shift at work, a social-media feed, a classroom, a short conversation, or a standardized test gives only a thin behavioral surface. From it we may learn something real about what a person is doing under those conditions. We do not thereby obtain the person's whole problem situation.

Epistemology does not require that everyone secretly be profound. That would be another unsupported inference. It requires a distinction between evidence of narrow current activity and evidence of principled epistemic inferiority.

The distinction matters because individuals differ not merely in what they know but in what becomes problematic to them. Two people can share language, education, workplace, and neighborhood while one notices injustice, another inefficiency, and another beauty. Those differences compound. Problems change knowledge; changed knowledge alters salience; salience generates new problems. Some are reinforced by social recognition, others suppressed.

The epistemic mind is therefore not simply a bundle of beliefs. A person is an evolving landscape of problems. The universal explainer, on this account, is not a person displaying continuous brilliance. The child asking why, the worker narrowed by an assigned task, and the person trapped in an unresolved tension are different visible states of an epistemic life.

That gives biography another dimension. Alongside institutions, relationships, achievements, traits, and events runs a history of which tensions became alive, which questions survived, which criteria changed, and which problems reorganized what came next. That history is not fully visible from outside, but neither is it sealed inside subjectivity. Problems can be externalized, reconstructed, criticized, and shared. Your mind remains solitary; your problem situation is not therefore private in principle.

Visible behavior therefore gives evidence about what a person is doing under particular problem situations. It does not automatically reveal the principled limits of the problems that person could ever understand or create. What you can explain tomorrow depends partly on problems you have not yet encountered, distinctions you have not yet learned, criticisms you have not yet received, and criteria that may not yet exist.

A present surface cannot contain that future.

That is enough to return to the objective problem of problems with a sharper question: what exactly can an externalized problem carry between minds, and what remains unexplained when it arrives?

8. When a problem becomes yours

A problem can become objective knowledge. It can be stated, preserved, criticized, taught, and inherited. Its criteria can become objective knowledge too. Later minds can reconstruct enough of the structure to continue the inquiry.

That may explain how a problem can travel. It does not explain why the problem becomes alive in you.

The written problem does not contain the whole knowledge history through which you encounter it. It does not contain which distinctions are already salient or which experiences make the tension matter. It does not contain the competing problems already active in you, or the inexplicit criteria through which something can feel wrong before you can say why.

Objective knowledge can therefore carry the structure of a problem without carrying a complete explanation of its uptake.

The personal cases make that limit visible. The lonely young man's problem can be stated objectively, but the statement alone does not explain why one reformulation reorganizes his life and another does not. A deadening job can assign an objective task without revealing the richer problem world that becomes visible after work. A standardized test can record performance on a defined problem without containing the person's complete problem history or establishing the limits of the problems they could later learn to form.

The same limit appears with novel criteria. A composition can embody a new musical possibility. A proof can embody a new standard of elegance. A philosophical distinction can become available to later readers. But the artifact does not yet explain how a solitary, path-dependent mind comes to recognize the new criterion as a criterion. As something that matters.

Objective knowledge is therefore not a rival explanation of you. It is part of your environment. It supplies problems, criticisms, distinctions, examples, and criteria that can be reconstructed. You bring an evolving problem situation through which those objects become differently salient. Their interaction changes both what you can see and what objective knowledge you can create next.

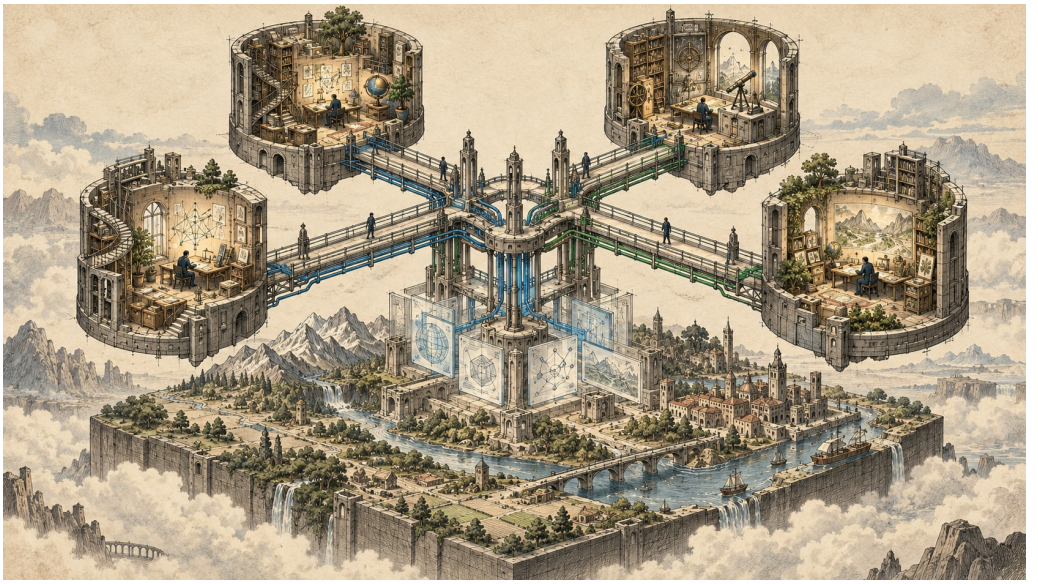
That explains why a shared culture need not produce identical minds. It does not explain salience itself.

The gap has therefore moved backward. Before a problem can be criticized, it must in some sense appear. Before a criterion can be revised, something must make the existing criterion inadequate. Before an interest can recur, some distinctions must become available to attention rather than others.

You may encounter the same gap here. Some aspects become salient, others do not. Some produce resistance; others reorganize a prior explanation. The process begins before its causes are fully available to introspection.

The problem of problems therefore creates a successor problem: How does anything become salient enough to become a problem at all?

X. Problems Aligning



You are an unsettled and unsettling individual.

That picture is exhilarating when the subject is freedom. It becomes frightening when the subject is coordination.

If every person is such a system, how do billions of them get anything done? How do they raise children, build bridges, run hospitals, trade with strangers, operate businesses, govern cities, form friendships, or simply agree to meet at eight?

The familiar answers arrive quickly: humans are social animals; civilization requires hierarchy; individuals sometimes have to yield to a larger purpose. Without enough authority or structure, the story goes, coordination gives way to selfishness, irrationality, or conflict.

These explanations are so familiar that they often arrive as conclusions before the problem has been made explicit.

Epistemology offers a different place to begin. The coordination problem should submit to the kind of solution already partially instantiated within your epistemic mind.

A mind is already a coordination problem of incalculable complexity. And somehow, imperfectly and mysteriously, it works.

1. Coordination Begins Inside a Person

You do not appear internally simple. A single mind contains memories, habits, skills, expectations, fears, interests, inherited explanations, inexplicit abstractions, unfinished problems, emotional pressures, and changing criteria. These structures do not all agree, become explicit at once, or point in the same direction.

You can want incompatible things and believe propositions that conflict. An old explanation can survive for years before a new observation makes its contradiction visible. You can also know what you ought to do while finding a different problem more compelling. Yet you normally experience something like a self.

That apparent unity is not yet explained. Calling it "the self" merely names the phenomenon if we imagine a little manager inside the mind coordinating everything else. We do not yet possess a causal account of how subjective unity, explanatory creativity, problem formation, and the physical processes of the brain fit together.

But incomplete is not empty.

Across the previous essays we developed several broad structures that appear to regulate the epistemic mind. Problems become salient. Tentative explanations arise. Criticism eliminates some errors. Reality resists arbitrary expectations. Explanations that survive one problem can collide with explanations used elsewhere. Internal consilience can produce a temporary equilibrium, and new contradictions can reopen it. Interests make some classes of problems recurrently available. Objective knowledge extends the reach of the biological system.

None of this requires a perfect internal controller. It describes a system that can remain directed by continually self-correcting.

That distinction changes the coordination problem. Order need not mean command. Coordination need not mean that every component has the same goal. Complexity need not imply that one privileged center has calculated the state of the whole.

Inside you, enormous epistemic heterogeneity somehow remains sufficiently coordinated for action to continue. The fact does not establish that societies work the same way. A society is not a nervous system, a firm is not a consciousness, and a government is not the frontal cortex of a nation. But the individual case weakens one familiar intuition: complexity by itself does not establish the need for a central controller.

The skeptic asks how billions of autonomous explainers could possibly avoid chaos. The first mystery is that one solitary mind does. Billions of solitary minds do. It doesn't begin to explain the coordination problem but it deepens it.

Reality helps explain why that internal coordination is not completely arbitrary. Your abstractions do not exist in a vacuum. Expectations fail. Bridges collapse if their structure is wrong. A tool does not work merely because its user prefers the theory

behind it. Other people behave differently from the models we construct of them. Bodies age. Resources have physical constraints.

Reality does not write our explanations for us. It constrains them by resisting some of what we conjecture. That resistance gives otherwise divergent abstractions something outside themselves against which they can fail.

This is not certainty, direct access to reality, or the old dream that enough facts will force all rational people to agree. The claim is weaker: there is one reality, and different explanations of it remain exposed to consequences they do not control.

That matters when we move from one mind to many. Different minds do not need identical internal landscapes to encounter the same bridge, contract, disease, customer, storm, or traffic signal. They are different explanatory systems acting in one world.

Reality supplies common constraints without supplying a common mind.

2. Bridges Without Merger

Shared reality is not enough. Two minds can face the same object and still understand it differently. Coordination requires some way for the products of one mind to alter the problem situation of another.

Communication does this without transmitting a mental state intact. One person constructs an expression. Another reconstructs something from it. The result in the second mind is never simply a copy of the first, but a copy is not required. A sentence can expose a contradiction, a drawing can reveal a hidden dependency, and a price can make a resource constraint visible. The second mind only needs to be perturbed in a way that changes what can be noticed, criticized, or attempted next.

Two minds do not need to merge in order to become mutually corrigible. The social process is real. The collective knower is not.

The same pattern extends beyond conversation. Solitary minds coordinate through structures that exist outside any one of them: language, writing, contracts, laws, standards, schedules, money, technical interfaces, scientific papers, software, databases, road signs, and calendars. These structures are objective knowledge. They can persist after their creators have left, be criticized by people their creators never met, and coordinate people who disagree about almost everything else.

A bridge does not require every engineer, steelworker, inspector, financier, truck driver, and future traveler to share the same internal problem world. The participants need compatible interfaces for their different contributions to meet.

The more people a coordination structure has to serve, the less it can depend on the full internal world of each participant. In that sense, coordination scales by becoming thinner. A traffic light need not know why each driver is traveling. A contract need not

contain the parties' philosophies of life. A technical standard can fix an interface while leaving the work above it open.

Objective knowledge therefore allows deep internal difference to coexist with narrow external compatibility. Society need not be one mind coordinating its parts. It can instead be a world dense with objective structures through which different minds become partially predictable to one another.

That solution contains its own danger. The interface works precisely because it represents only part of the person.

3. Compression and the Appeal of Authority

Institutions cannot represent the full epistemic life of every person participating in them. They need roles, categories, schedules, procedures, credentials, permissions, responsibilities, and expected outputs. A workplace sees an employee, a school sees a student, a court sees a litigant, and a bureaucracy sees a case.

These abstractions are not necessarily mistakes. They can be useful solutions to real coordination problems. The mistake begins when the compression is mistaken for the person: the role becomes an ontology, the category becomes an explanation, and the hierarchy becomes evidence of human rank. A technology for coordinating people starts to make claims about what those people are.

This is one place where the authority memeplex reappears. Its attraction need not begin as a moral theory. Hierarchy is cheap.

If ten people must negotiate every decision with every other person, coordination becomes costly. If one person decides and nine comply, the decision path becomes simpler. Authority can reduce uncertainty, assign roles, fix priorities, and permit action before deep shared understanding exists.

That is a real solution to a real problem, which helps explain why hierarchy appears so persistently even in cultures that celebrate liberty. Workplaces assign problems and outputs. Schools impose curricula. Families encode responsibilities and rank. Games deliberately construct rules, penalties, referees, captains, and winners.

None of this proves that human beings secretly crave subordination. It may instead show that hierarchy has long been one of the cheapest available coordination technologies.

A useful technology can then harden into a theory of human nature. "Someone has to be in charge" moves from a practical conjecture to an assumed law. "People need structure" becomes "people cannot be trusted with freedom." A manager's local knowledge becomes evidence of epistemic superiority. An adult's responsibility becomes evidence that a child is not yet an explainer. A ruler's need to act becomes evidence that a population needs ruling.

The transition is easy to miss because the original coordination problem is real. The error is not that coordination requires constraints. It is that a successful local constraint is promoted into an explanation of the person being constrained.

4. Liberty, Constraint, and Process

The Enlightenment becomes relevant to the argument. It matters here neither as a sacred historical episode nor as an authority to which the epistemology must conform. The relevant point is that some liberal institutions can be described in terms already supplied by the epistemology.

If knowledge grows through conjecture and criticism, then institutions that preserve criticism preserve a condition of error correction. If different people create different problems, then pluralism preserves variation that uniformity would remove. If explanations are fallible, then structures that keep authority criticizable preserve the possibility of revising the rules.

Free speech can therefore function as an enabling technology for criticism. Free association permits recombination among different problem worlds. Freedom of conscience leaves room for your internal criticism to continue beyond an externally preferred conclusion. Exit lets you leave a problem environment that has become closed to revision.

These institutions remain morally interpretable in many ways. The present argument does not need to settle those theories. Its narrower claim is explanatory: some liberal forms appear consilient with conditions under which fallible explanatory systems can continue producing and correcting knowledge.

Perhaps the Enlightenment discovered parts of a social epistemology before we possessed an explanation of why those discoveries worked.

That formulation immediately creates a problem. Political liberty can coexist with highly constrained epistemic life. You may possess the vote and freedom of speech while spending most of your waking hours inside someone else's problem formulation. You may be free to criticize the government while finding that questioning a workflow threatens your employment. You may choose an institution freely and still find little room to challenge its assumptions once inside.

The point is not that political liberty is meaningless. It is that political principles do not by themselves solve the coordination problem inside ordinary institutions.

Freedom at scale still requires constraints. A programming language restricts syntax and thereby makes a shared technical space possible. A road protocol constrains drivers so that many independent journeys can coexist. A technical standard fixes an interface so that variation can proliferate above it.

The useful distinction is therefore not between constraint and freedom. It is between constraints that preserve or enlarge future problem spaces and constraints that close them. A coordination structure can specify what must be shared while leaving unrelated questions open; it can also expand its jurisdiction until the interface is treated as the person.

That replaces the vague instruction to "balance freedom and order" with a more specific question:

> Which constraints make more independent epistemic lives possible?

The answer can differ across a constitution, a family rule, a workplace process, a school curriculum, a software protocol, or a market institution. The criterion remains the same even when the local tradeoff does not.

The criterion becomes harder when coordination is economic, because one person's open problem space does not create an automatic claim on another person's resources.

Suppose you develop a deep interest in an obscure form of music, an abandoned language, or an unfashionable mathematical problem. There may be no market for it. An employer has customers, deadlines, payroll, quality requirements, contracts, equipment, and risk. The employer can reasonably say:

> Your interests are your problem. If you want me to pay you, I need you to work inside this process. That is my problem situation.

The employer is also a solitary problem-selector. The worker's interest does not establish an economic obligation on the employer, while the employer's need for predictable output does not establish epistemic superiority over the worker. They are different problem situations looking for a structure of exchange.

Economic institutions can sometimes provide such an interface. A buyer and seller need not agree about why the exchange matters. A worker may want security while an employer wants reliable output. Exchange can coordinate those purposes without replacing them with one shared goal. Free association permits new combinations of people and knowledge. Entrepreneurship lets someone act on a problem others have not yet recognized. Competition exposes organized solutions to criticism by alternatives.

Some features of economic liberty therefore resemble the epistemic properties already identified in political liberty: variation, association, experimentation, criticism, and exit. That does not amount to a theory of liberal economics. Markets can concentrate power. Exit can be costly. Economic value and epistemic value are not the same. Firms can create epistemic confinement.

The objection does not disappear. It becomes more precise.

Return to your loved one. At work, they appear reduced because the process may already specify the problem, what counts as a solution, the sequence of action, and

even the acceptable language for describing failure. The worker still supplies attention, memory, judgment, and action, but much of the epistemic landscape has been closed in advance.

There can be good reasons for doing this. Variation can be expensive or dangerous. A factory cannot allow every worker to redesign the product on the line. An airline cannot let each pilot invent a new landing protocol. A hospital cannot make every safety procedure optional. Standardization can encode hard-won objective knowledge.

So the explanatory problem is not whether process is bad. It is how much of a person's problem-generating capacity must be constrained to solve the actual coordination problem. A second question then appears: which remaining constraints exist only because the available coordination technology cannot cheaply adapt?

Historically, the answer has been: quite a lot. Hierarchy and standardized process were available. Deeply adaptive coordination was not. Some forms of social and economic constraint may therefore have survived for a different reason. The people inside them may have been capable of greater autonomy while cheaper ways to coordinate heterogeneous minds were unavailable.

That conjecture changes the criticism. "Democracy is failing" and "capitalism is failing" remain possible problem statements, but neither phrase yet identifies a mechanism. Failing how? Which part of the tentative solution is implicated? What observation would distinguish a failure of liberty from a failure of the institutions through which liberty is being coordinated?

Polarization, for example, could implicate free speech, information architectures, political incentives, or some interaction among them. Alienating work could implicate voluntary exchange, the practical cost of exit, primitive coordination technology, concentrated authority, or something else. The labels do not decide among those conjectures.

The epistemic stance is therefore not to protect democracy or capitalism from criticism. It is to keep the criticism open until a mechanism has been proposed. Liberal institutions remain tentative solutions inside the same cycle as everything else: problem, tentative solution, error elimination, new problem.

No institution becomes exempt from criticism because it is liberal. No criticism becomes explanatory merely because the institution disappoints us.

5. Society Is Not the Problem

The argument returns to you.

We routinely say that society is losing faith in democracy, the public thinks capitalism is broken, the market wants growth, or the government knows what must be done. These phrases are useful compressions. Taken literally, however, they recreate the collective mind we have spent several essays dissolving.

There is no social subject contemplating the failures of liberalism. There are people reconstructing objective structures inside different problem situations.

A newspaper headline, political meme, economic theory, institutional rule, or workplace procedure can exist outside any one person. But criticism becomes epistemically active only when someone reconstructs one of those structures as a problem. "Democracy is failing" can circulate through millions of artifacts and mouths without becoming the same problem twice. One person may reconstruct it as a problem of representation, another as a problem of information, and another as a problem of authority. Someone else may not experience it as a live problem at all.

The same is true inside a workplace, family, market, or school. A rule, role, process, or status structure can be objective knowledge. The tension between that structure and your other explanations occurs inside your particular mind with your particular history of interests, expectations, problems, and criticism.

That is where coordination reaches you.

The institution never receives your whole epistemic system. It receives behavior through an interface. You never receive the institution as a whole either. You reconstruct fragments of rules, reasons, expectations, incentives, stories, and consequences inside your own problem situation.

Coordination therefore does not consist of a collective subject solving one problem. It consists of solitary problem-holders becoming partially coupled through objective structures.

This changes what a coordination failure can be. The shared interface may be too thin for the task. It may encode bad knowledge. One side may be unable to reconstruct the constraints faced by the other. A role may suppress criticism the system actually needs. Exit may be formally available but practically inaccessible. Or a compressed representation of you may have been mistaken for you and used to close problems that remain alive for you.

None of those explanations requires a social mind. They require people, objective knowledge, and mechanisms through which each changes the problem situations of the others.

The point becomes personal because the social world does not revise itself above you. A criticism becomes live when you reconstruct it. A rule becomes a problem when it collides with something else you know. A role becomes confining when the interface closes a problem space you are trying to open. A shared project becomes possible when enough objective structure lets you act with people whose inner landscapes you cannot inspect.

Society is not making up its hive mind. You are reconstructing your part of the social world inside your own mind from objective structures that other people have made available for reconstruction and criticism.

That is precisely why the next technology frontier matters.

6. AI Changes the Coordination Technology

Much contemporary anxiety about artificial general intelligence imagines a future entity that is difficult to control. It can form goals, revise strategies, outgrow its creators, resist instructions, and reorganize the world around problems we did not choose.

The substrate would be new. The coordination fear is not.

We already live among general intelligences that change goals, reject authority, create new problems, outgrow their teachers, reinterpret instructions, develop unexpected interests, and sometimes refuse to cooperate. They are biological general intelligences: people.

Before AGI, there was BGI.

The analogy has limits. A hypothetical artificial person might operate at different speeds, copy itself, consume different resources, or possess capabilities no biological person has. Those differences could create new coordination problems. The narrower resemblance is the problem of autonomy: how do we coordinate something that cannot be perfectly aligned to our goals?

Civilization has spent millennia constructing partial answers through laws, norms, markets, rights, schools, hierarchies, contracts, constitutions, bureaucracies, and systems of surveillance. The AGI alignment problem therefore enters a species with a long history of trying to align generally intelligent entities.

That makes another distinction important. Contemporary AI can exceed an unaided person's capabilities in particular tasks. It can search and compare large bodies of text rapidly, manipulate symbols, translate languages, calculate, and synthesize across domains. Those capacities do not by themselves establish personhood.

Under this ontology, the threshold would be different. It would concern autonomous creation of new explanatory knowledge, self-originating problem formation, criticism, revision of evaluative criteria, and the other structures associated with persons. Current AI can execute methodologies, recombine objective knowledge, and participate functionally in explanatory work without thereby becoming an autonomous epistemic subject.

The more immediate social configuration therefore does not require AGI. It is already intelligible as:

> billions of biological general intelligences with access to unusually powerful objective-knowledge tools.

A book stores objective knowledge. A database structures it. A search engine retrieves fragments of it. An AI system can add a further operation. It can reorganize available

objective knowledge around the current problem posed by an individual, then reorganize it again when criticism changes that problem.

That sequence matters. You prompt the system from inside your idiosyncratic problem situation. The system produces a reconstruction that crosses oceans of objective knowledge not originally organized around that problem. You criticize the reconstruction. The criticism changes what now needs explaining. The objective landscape can then be reorganized around the changed problem.

Human teachers, librarians, editors, consultants, friends, and experts have always adapted knowledge for other minds. The difference here is not absolute kind but scale, speed, availability, and breadth. And the incumbent authorities are understandably anxious.

The result is a new practical interface between a solitary mind and vast objective knowledge. It does not make the AI a person. It changes forever what a person can reach.

Previously, access to AI could be framed mainly as an individual extension. Equipped with better tools, you can compare more, remember more, and reach further into objective knowledge. The coordination problem changes the question. Social coordination depends on objective structures that let different minds partially reconstruct one another's constraints. A technology that adapts those structures to different problem situations can therefore alter coordination itself.

A team may be divided by incompatible vocabularies. A worker may see a procedure without understanding the business constraint behind it. A citizen may receive a policy argument compressed into partisan language. In each case, the obstacle may not be disagreement alone. It may be that one problem landscape is poorly represented inside another.

AI can lower that translation cost. It need not merge the minds. Of course it can't. But it can help each person reconstruct enough of the other's objective structures to formulate a better problem.

That is the important shift. The technology may matter socially not only because it makes individuals more capable, but because it may help heterogeneous groups communicate and coordinate without requiring their members to become more alike.

The same capability can move in the opposite direction. AI can increase surveillance, make people more legible to institutions, personalize persuasion, enforce procedures, or concentrate objective knowledge behind proprietary interfaces. Nothing about the technology guarantees liberty.

The symmetry is the point. The technology is morally neutral. A tool that lowers the cost of coordinating autonomous people can also lower the cost of controlling them.

Historically, hierarchy often had an advantage because it was cheap. Dynamically adaptable objective knowledge may lower the cost of coordination among people who

retain independent problem situations. If it does, some apparent tradeoffs between freedom and order may not be permanent features of human nature. They may be temporary features of our now antiquated coordination technologies.

That is an optimistic conjecture, not a conclusion. It opens a new problem space.

7. You Are Inside the Experiment

This argument is recursive; this essay was produced through the kind of coordination it describes.

A person entered the inquiry with a problem. Call him the **operator**. An artificial system preserved context, retrieved and compared formulations, extrapolated patterns, and generated candidate reconstructions. But those operations did not fix the problem in advance. Objections changed what counted as a good answer. Some formulations were rejected. New distinctions appeared. The operator changed the resolution of the inquiry, and the system reorganized objective knowledge, including the operator's insights, around the changed problem.

No shared consciousness appeared. No hive mind was required. A solitary human mind coordinated with objective knowledge through an artificial system.

Now the result reaches another solitary mind: yours.

You do not receive the operator's mental state. You reconstruct this argument inside your own epistemic landscape. Some sections may become salient because they collide with experiences or explanations the author never possessed. Some may appear obvious, others wrong, and still others may create a problem that did not exist for you before reading them.

If you take one of those tensions into your own AI-mediated inquiry, you become the operator, and the knowledge evolves. The text is objective knowledge. Your criticism changes the problem. The available objective landscape is reorganized around your new problem.

A social epistemic process has then occurred without a social mind. This is not merely an analogy to the coordination mechanism under discussion. It is an instance of it.

That instance also reveals a capability asymmetry. Institutions, governments, companies, experts, competitors, and individuals can all use tools that reorganize objective knowledge around a target problem. If such tools materially expand explanatory reach, people without comparable access will be less able to reconstruct the problems in which they are being asked to participate.

The deeper opportunity is not epistemic warfare. It is participation at a different scale. You may once have had to defer because relevant knowledge was dispersed across inaccessible fields. You now may be able to reconstruct enough of the problem to criticize it productively, if that's where your interests pull you. A worker may understand an

employer's constraint. An employer may see where a process unnecessarily crushes variation. A citizen may reconstruct the strongest explanation behind a policy rather than responding only to its slogan.

"I may be wrong and you may be right, and by an effort, we may get nearer to the truth." It's a beautiful compression, but effort can be crushing. AI cannot eliminate the effort, but it can reduce the coordination costs substantially.

8. The Greater Good Without a Greater Mind

The language of the greater good often appears where coordination becomes difficult. Billions of problem situations is overwhelming, so the complexity is compressed into a superordinate problem: one goal, one metric, one direction. Individual interests are then evaluated through that model.

Sometimes common constraints are real and urgent. A contagious disease, a military attack, a failing power grid, or a collapsing bridge can create problems around which many people have to coordinate quickly. Powerful AI systems operating uncontrollably would certainly meet that standard.

But the "greater good" is not a mind. It is an objective model of goods, harms, priorities, and tradeoffs. Someone conjectured it. Different people reconstruct it. Someone can criticize it.

That does not make common action impossible. It makes the epistemic structure explicit. The question is not whether individuals ever accept constraints for common purposes; they plainly do. The question is what happens when a collective purpose is treated as if it could no longer be criticized as a problem.

Freedom preserves an open future. Openness has an intimate form: the capacity for new problems to become alive in a particular person. The coordination problem now extends the same idea into groups.

Individual flourishing does not require escape from society. It requires social structures in which belonging does not demand epistemic flattening. A group can have rules, responsibilities, shared projects, and even hierarchy where the local problem requires it. The relevant distinction is whether those structures remain criticizable. In the form being described here, members retain the capacity to form problems the group did not assign, and the shared interface remains narrower than the person who participates through it. Authority then remains a coordination relation rather than an ontology of superior and inferior minds.

This is not a blueprint for an ideal society. It is an explanatory criterion: does this structure preserve more independent error-correcting lives, or does it unnecessarily close them?

That question scales from two people to billions.

Seen this way, liberty is not a completed historical discovery. The Enlightenment did not complete the problem of human coordination. Democracy, capitalism, free speech, and rights did not complete it either. They remain tentative solutions whose failures can reveal what their underlying assumptions missed.

A liberal institution may preserve criticism in one layer while centralizing it in another. A market may formally permit exit while making it practically expensive. Speech may be formally free while an information architecture rewards low-quality conflict. An institution designed to protect liberty may still depend on workplaces, schools, families, or bureaucracies whose coordination logic was inherited from older authority structures.

The same epistemic stance that makes liberty intelligible therefore prevents liberty from becoming dogma. The old explanations remain open to the process that produced them: problem, tentative solution, criticism, and a new problem.

That is creative criticism.

The explanation remains incomplete. We still do not know how one mind achieves its apparent unity. We do not know the full conditions under which free institutions remain stable. We do not know whether AI-mediated coordination will predominantly enlarge individual problem spaces or make individuals more governable. We do not know what forms of artificial personhood, if any, may eventually appear.

The incompleteness does not erase the direction of the argument. The coordination problem does not require a mind above minds. It may require better bridges among them.

If that is right, some apparent conflict between freedom and order may have another source. It may reflect limitations in the objective knowledge through which people have been coordinated. It would require better explanations and better technologies for allowing people to remain epistemic.

You may already be testing the mechanism while reading this. You are reconstructing the argument inside a problem situation no person can fully know, perhaps with an AI beside you reorganizing the same objective knowledge around your own questions.

That success illuminates the next problem.

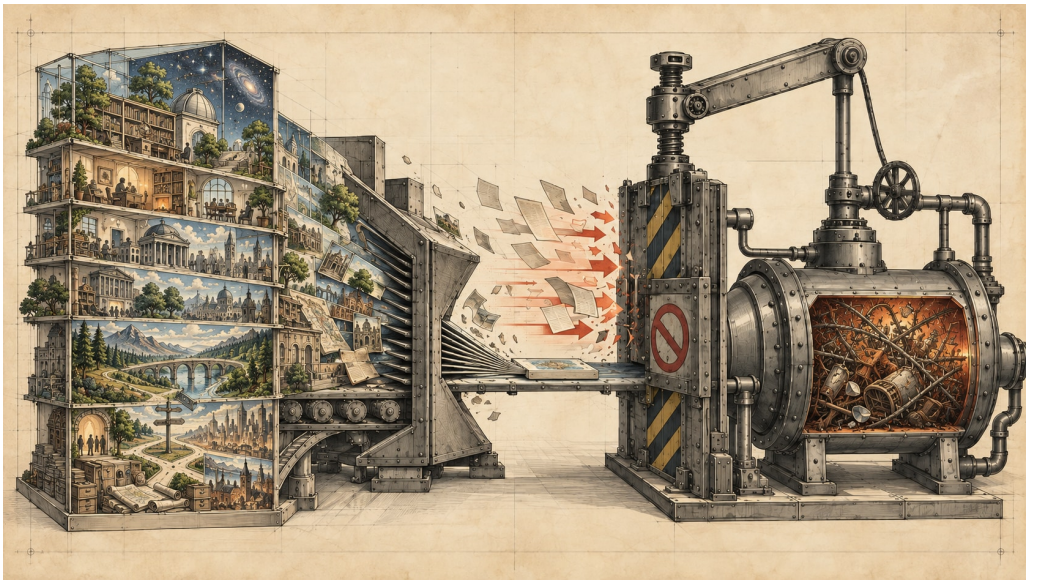
The same technology that can equip you can equip people that want to understand, predict, persuade, rank, and control you. Dynamic objective knowledge can widen your problem situation. It can also make you compressible and easier to control.

The question is no longer simply whether billions of minds can coordinate. It is whether the technologies that make deeper freedom possible can remain aligned with the freedom they enable.



The Unfinished Explainer

XI. Becoming Evil



Perhaps the problem of coordinating billions of free explainers was never simply a conflict between liberty and order. Perhaps some of the tradeoffs we inherited reflected the limits of our coordination technologies. Perhaps objective knowledge could increasingly carry the burden once assigned to hierarchy, standardization, and control.

That possibility creates a darker objection. What if the problem is not coordination? What if the problem is us?

Perhaps instead people require control because human nature is dangerous. Perhaps our ideas can become transcendent while our instincts remain primitive. Perhaps we repeatedly build freer institutions only to corrupt them. Perhaps the same biological organism that writes constitutions also wages war, humiliates strangers, seeks dominance, hoards status, rationalizes cruelty, and turns difference into hierarchy.

This is one of the oldest explanations of the human condition. It appears in theology, politics, moral philosophy, psychology, and cognitive science. The vocabulary changes. The structure often does not. Human beings are flawed in some deep way. Their nature cannot be trusted. Therefore order must come from somewhere else.

God. Law. Tradition. The ruler. The priesthood. The expert. The institution. The algorithm. The manager. The authority memeplex has always had a moral core. It does not merely say that coordination is difficult. It says that people are dangerous and require control.

Not whether people do terrible things. They do. Not whether suffering is real. It is. Not whether moral responsibility matters. It does.

The question is deeper:

> What is the explanation is evil we're carrying?

And if your exceptionality is epistemic, what would an explanation of evil look like if it began there?

1. Evil Is an Explanatory Problem

The word *evil* does more than classify unpleasant behavior. It performs explanatory work.

It tries to tell us why suffering occurs, why some actions are morally intolerable, why some people deserve blame, and why social order requires restraint. That is why explanations of evil do not remain inside philosophy. They generate institutions.

If evil is explained as demonic corruption, spiritual authority becomes relevant. If it is explained as fallen human nature, obedience to divine or moral law becomes necessary. If it is explained as irrational instinct, experts and institutions gain legitimacy as regulators of defective cognition. If it is explained as social deviance, conformity becomes morally charged. If it is explained as innate aggression, coercive control can begin to look realistic rather than tragic.

The explanation of evil helps determine the explanation of governance.

So the first question is not:

> Who is evil?

It is:

> What explanatory work does the concept of evil perform?

Traditional moral cosmologies often answered by placing moral order in a larger metaphysical structure. Sin. Divine law. Punishment. Redemption. Heaven. Hell. Cosmic justice.

These ideas addressed a real problem. The world often does not appear morally balanced. Cruel people prosper. Kind people suffer. Justice is incomplete. A moral cosmology can restore order by moving the final accounting beyond ordinary life. Reward can be deferred. Punishment can become deserved. The authority of the moral system can be located outside the individual.

That does not make every theological explanation crude. Many traditions contain sophisticated accounts of guilt, temptation, self-deception, repentance, conscience, forgiveness, and transformation.

The epistemic problem is narrower.

When the source or resolution of evil is located primarily in an authority outside the person, the explanation can leave obscure the mechanism by which moral knowledge is preserved, criticized, or revised in the person who actually lives with it.

You suffer here. You form explanations here. You reconstruct inherited rules here. You feel guilt here. You criticize a doctrine here. You change here.

A moral cosmology may tell you where the account is finally settled without yet explaining how the moral structure operating in you was produced. Why did you preserve one harmful explanation and revise another? Why did suffering remain trapped? Why did criticism fail? Why did a rule become part of your problem situation at all?

Hell makes the problem especially sharp if we interpret it as permanent moral finalization. No further revision. No successor problem. No possibility of becoming less wrong. No reopening of the moral landscape.

From a fallibilist epistemology, that is not merely punishment. It is the destruction of corrigibility. The person is fixed. The explanation of who they are becomes eternal.

The severity is therefore structural as well as painful. A fallible person has been converted into a finalized moral object. Within an epistemology built around open-ended error correction, permanent moral finalization is not an extension of personhood. It is the cancellation of the process by which a person can make progress.

That already tells us something important. Harm alone cannot be the whole explanation of moral evil. A lion kills. A scorpion stings. A disease destroys tissue. A storm kills people. We do not ordinarily call these things evil in the same way we call deliberate cruelty evil.

Why not?

Because the moral category becomes intelligible only where alternatives, explanation, criticism, and responsibility are meaningful. A lion does not stand before competing moral explanations of predation. A scorpion does not revise its theory of justice.

If the distinctive property of persons is epistemic, then a theory of human evil grounded only in some non-epistemic animal essence becomes unstable. Aggression, territoriality, impulsiveness, tribal behavior, and status seeking may identify real patterns. But those patterns alone do not explain moral evil. Otherwise the lion would be wicked and the scorpion corrupt.

The moral difference lies in the existence of a creature capable of explanation.

That does not mean evil is merely ignorance. You can know many facts and still be cruel. A tyrant may understand suffering very well. A manipulator can be strategically brilliant. Scientific knowledge does not prevent rationalized harm.

The relevant failure concerns explanatory knowledge: the model of another person, the problem you think you are solving, the criteria by which suffering counts, the alternatives you can imagine, and the criticisms your explanation can survive.

You may lack a better explanation of someone else. You may lack a way to coordinate without domination. You may preserve a conception of yourself that requires another person's humiliation. You may not yet possess the criticism that can penetrate an identity-bound explanation. You may continue solving a problem whose evaluative criteria are themselves producing the harm.

So the epistemic conjecture is not that evil is ignorance. Moral evil becomes possible when harmful explanations become closed to criticism and thereby preserve avoidable suffering among corrigible persons.

The distinctions matter. The harmful explanation is the content. Blocked criticism is the mechanism of closure. Persistent avoidable suffering is the morally relevant outcome.

That is not a moral theory. It is a conjecture about where the explanatory problem belongs.

2. How Harmful Explanations Become Closed

The difficult implication is that moral harm does not arise only when reason fails. Reason is conjectural.

You can construct an elaborate explanation of why cruelty is necessary. A society can develop sophisticated objective knowledge that stabilizes domination. An institution can create a technically precise system for treating some people as less important than others. A moral theory can turn punishment into virtue.

The danger is not simply the beast beneath reason. Explanatory sophistication does not imply moral correctness. The same capacity that lets you invent a better bridge can let you invent a better prison. The same abstraction that lets you universalize dignity can let you universalize hierarchy. The same skill of consilience that makes a worldview powerful can make a harmful worldview internally coherent.

Fallibility therefore sits inside explanation, not outside it. The question is whether the explanations governing action remain open to criticism.

Imagine someone you dislike. Perhaps they are arrogant. Perhaps they betrayed someone. Perhaps they humiliated you. Perhaps they hold a political view you find repellent.

You form an explanation. They are selfish. Cruel. Stupid. Narcissistic. Dangerous. Perhaps part of the judgment is accurate.

But then the explanation begins to control what becomes visible. A kind act looks manipulative. A mistake looks intentional. A disagreement confirms corruption. A new fact is reinterpreted as evidence for the same model. The other person becomes thinner as your explanation becomes more stable.

The surface becomes the person. The category becomes the explanation. The model becomes self-preserving. At that point moral judgment can become epistemic closure. You no longer need to understand the other person because you already know what they are.

This is one way evil can begin without drama. Not as a monstrous impulse but as a narrowing explanation.

The moral consequence of a thin model is not merely inaccurate description. It changes what actions appear permissible. A coworker may appear dull because work exposes only a narrow slice of their problem situation. A lonely person may appear defective because a population category has replaced the specific structure of their life. A child may look irrational because the adult cannot see the problem organizing the child's attention.

Once epistemic depth disappears from the model, criticism becomes less necessary. If the employee is merely lazy, control looks justified. If the child is merely immature, command looks justified. If the dissenter is merely dangerous, suppression looks justified. If the out-group is merely irrational, exclusion looks justified.

Abstract it.

The structure is roughly:

$$\text{thin model} \rightarrow$$
$$\text{reduced perceived epistemic depth} \rightarrow$$
$$\text{reduced need for criticism} \rightarrow$$
$$\text{greater tolerance for coercion.}$$

That is one way the authority memeplex acquires moral force. But the explanation can go wrong even earlier.

Consider two people in conflict. One asks:

> How do I make them obey?

The other asks:

> How do I preserve my dignity?

The first person may solve their problem competently. They may improve incentives, design stronger penalties, optimize enforcement, and become increasingly effective. The harm can increase even as the local solution improves because the problem formulation itself is wrong.

The same structure appears in families, workplaces, governments, schools, and intimate relationships.

A parent asks, "How do I stop this behavior?" instead of, "What problem is my child trying to solve?" A manager asks, "How do I enforce compliance?" instead of, "Why does this process repeatedly generate resistance?" A citizen asks, "How do we silence dangerous ideas?" instead of, "What makes these ideas persuasive?"

The epistemic cycle can operate perfectly inside a bad problem.

That is why criticism must eventually reach not only the proposed solution but the problem being solved. Everyone is becoming wrong all the time. That cannot be the moral distinction. You carry bad explanations. You misread other people. You may resist criticism. If evil were simply error, morality would collapse into ordinary fallibility.

The morally significant pattern appears when harmful explanatory structures become persistent and insulated from correction.

Abstract it.

Where the error is imposed on other people, one recurrent structure is:

$$\text{harmful explanation} + \text{power} + \text{blocked criticism} \rightarrow \text{persistent avoidable suffering.}$$

Power lets an error impose costs beyond the person who carries it. Blocked criticism prevents the explanation from becoming a successor problem. Persistence turns a corrigible mistake into a stable environment.

Evil therefore begins to look less like a substance and more like a process by which harmful explanations become morally consequential through closure.

That gives criticism a moral center. Free speech is not morally important only because people enjoy expression. Dissent is not valuable only because pluralism feels tolerant. Exit is not important only because autonomy is pleasant. These practices preserve pathways by which harmful explanations can become problems.

So a useful compression is:

> Preserve corrigibility where persons bear the cost of being wrong.

This is not the whole of morality of course. A corrigible person can still do harm. An open institution can still fail. But where people bear the cost of error, preserving the possibility of discovering and correcting that error becomes morally serious.

An authority that cannot be criticized can preserve a moral error indefinitely.

3. From Cognitive Error to Human Nature

The same inference can reappear without theological language. One modern route begins with observations from cognitive science. Humans anchor. Humans forget. Humans selectively attend. Humans seek confirmation. Humans rely on heuristics. Humans show framing effects. Humans reason differently under different conditions.

These observations can be valuable. For the present inquiry, the problem begins one step later: when a measured performance regularity is treated as an explanation of the person who exhibits it.

The move is subtle:

people often show confirmation bias

becomes:

people are confirmation-seeking creatures.

Or:

working memory is limited

becomes:

human reasoning is fundamentally bounded by that limit.

Or:

people use fast heuristics

becomes:

human nature is governed by irrational autopilot.

The empirical pattern may be real. The transfer to ontology is not automatic. An error pattern is not yet an explanation of personhood.

The popular autopilot picture is attractive because it appears to explain why people make mistakes: cognition is mostly automatic, shortcut-driven, evolutionarily biased, and only partially supervised by a narrow rational layer.

The inference then has a moral consequence. If the performance pattern is taken to reveal what a person fundamentally is, it can feed directly into the authority memeplex. If most cognition is irrational autopilot, serious freedom begins to look risky. People need nudges. Experts need to design environments. Institutions must compensate for defective judgment. Algorithms may need to protect people from themselves.

Some such interventions may solve real problems. The epistemic criticism is narrower: recurring failure under particular problem conditions does not establish a ceiling on explanatory universality.

You are not exhausted by your unaided biological processor. You have limited memory, so you write. You cannot inspect every relevant variable at once, so you use diagrams, models, software, checklists, statistics, and teams. You cannot remember an entire library, so libraries exist.

Compress that idea. The effective epistemic system has always been:

person + objective knowledge + tools.

A current biological limitation may become nearly irrelevant once objective knowledge externalizes it. That does not mean all limits disappear. It means a performance limit is not the same thing as a principled limit on explanatory possibility. To define human nature by unaided cognitive performance is like defining your mathematical capacity by what you can calculate in your head without paper. It confuses the substrate with the epistemic system.

The same problem appears in the distinction between fast and slow thinking. Some judgments arrive immediately. Others require effort. Some intuitive responses are wrong. Some deliberative processes are better. But knowledge changes the relationship.

A mathematician can see instantly what once required pages of work. A musician can hear a structural error immediately. A programmer can recognize a bug pattern before fully verbalizing it. An experienced craftsperson perceives possibilities invisible to a novice.

What was once slow can become fast. What begins as a fast impression can be unpacked into slow criticism. With tools, the system can become both at once: fast retrieval, slow evaluation, fast recognition, externalized memory, deliberate revision.

You are not stuck with one cognitive tempo. You learn. You compress. You externalize. You build tools.

Cognitive bias becomes more interesting, then, when treated as a problem rather than a verdict. If people repeatedly make a certain error, the observation creates an opportunity. Double-blind procedures. Randomization. Statistical methods. Audit trails. Checklists. Adversarial review. Pre-registration.

These are forms of objective knowledge created in response to recurring human error:

recurrent error → problem → method → reduced error.

The bias did not reveal a hopeless human essence. There's no such thing. It generated knowledge.

A pessimistic interpretation stops halfway through the cycle:

> Humans make errors. Therefore humans are defective.

The epistemic interpretation continues:

> Humans make errors. Therefore humans can create better error-correction structures.

The same observed failure can therefore support two very different explanatory pictures. One converts a performance regularity into a claim about human nature and

then uses that claim to justify stronger external control. The other treats the regularity as a problem for which new error-correcting knowledge can be created.

The first closes the explanatory picture of you. The second leaves it open to correction.

4. When the Structure Is Reconstructed in You

Cognitive science returns us to the personal problem. If harmful explanatory closure is possible wherever explanations are created and preserved, it cannot belong only to villains. The relevant machinery exists in you.

You can inherit a harmful explanation. You can create one. You can rationalize. You can reduce another person to a category. You can resist criticism when an explanation becomes identity-bearing. You can preserve a bad problem because changing it would destabilize too much else.

This does not mean everyone is equally evil.

Of course not.

It means the possibility is universal while the trajectory is individual. A child can inherit a hierarchy of worth. A family can teach that punishment proves love. A community can teach that outsiders are dangerous. A profession can teach that status indicates intelligence. A culture can teach that obedience indicates maturity.

These structures are not merely outside you. You reconstruct them. They connect to fear, identity, memory, status, affection, resentment, and other explanations already present in your problem situation. Some survive criticism. Some collapse. Some become deeper. Some are replaced by better models.

The path is never identical. Evil evolves in the solitary mind.

To see why, distinguish explanatory content from explanatory structure.

You may carry a harmful explanation such as:

> I am worthless.
>
> They are dangerous.
>
> Failure proves weakness.
>
> Control is the only way to remain safe.

Each can matter on its own. But persistent suffering often depends on how explanations are related.

Suppose you carry all three of these:

- I must never fail.
- Failure is inevitable.

- My worth depends on success.

The suffering does not belong to any one proposition. It emerges from the architecture. Each explanation reinforces the others. The problem cannot move. The system becomes locally self-preserving.

That is what *structural evil* or the *architecture of evil* means epistemically: not a demon inside the mind and not a second person, but a relation among epistemic states that helps preserve suffering by preventing the problem from moving.

This is also why criticism is not necessarily productive in itself. You can criticize yourself constantly and remain trapped. Rumination can look like relentless error correction while producing no successor problem. You return to the same accusation, the same evidence, the same failed solution, the same self-model:

$$P \to TS_1 \to EE \to TS_2 \to EE \to TS_1.$$

There is movement without development. Criticism has become captive to the problem formulation because every attempted correction is evaluated inside the same closed problem.

Flourishing therefore cannot mean maximum criticism. It requires criticism capable of changing the problem, its criteria, or the explanatory relations that keep reproducing it.

The same closure can appear in self-conception.

You may believe:

> I must be exceptional.

Or:

> I am permanently inferior.

Or:

> I must never need help.

Once the model becomes self-sealing, it determines what evidence is allowed to count. A compliment is dismissed. A setback is allowed to confirm the inexplicit theory. A success raises the standard. A relationship conflict becomes "proof" of unlovability.

You experience the consequences as reality because the explanation helps determine which evidence becomes salient, which alternatives remain thinkable, and which successor problems can appear.

The same can happen to an interest. Interests often support flourishing by drawing you into open-ended problem spaces. They create depth and generate knowledge. But

an interest can become closed, compulsive, identity-bound, and detached from criticism. More and more of life can be organized around a problem whose evaluative criteria are harming you.

The point is not to moralize unusual interests. It is structural. Any problem world can become narrow enough that you can no longer revise it. The same epistemic machinery that makes fascination generative can make fixation destructive.

How can billions of biological general intelligences ever coordinate? Here, a microcosm of that problem is revealed. Inside you, one model can dominate another. One problem formulation can suppress alternatives. One inherited authority can silence criticism. One self-conception can punish every deviation. One interest can crowd out the rest of the problem landscape.

The external moral story often imagines:

> evil actor → victim.

Inside one person, the architecture can be more complicated. The explanatory structure that inflicts harm and the structure that suffers from it can coexist in the same solitary mind—not as literal sub-persons, but as relations among often inexplicit knowledge.

Your explanatory landscape can therefore contain both a structure that constrains and the suffering produced by that constraint.

This point needs a guardrail.

If suffering becomes instantiated in your explanatory structure, that does not mean you caused it. Location of manifestation is not location of blame. An abusive environment can help produce a damaging self-model. A coercive institution can narrow problem space. Poverty can eliminate exit. Trauma can reorganize salience. A family can make criticism dangerous. A culture can make shame the default explanation of failure.

The fact that such structures later operate inside you does not transfer responsibility from their external causes to you. You may carry harmful objective knowledge you never chose and cannot yet articulate.

The epistemic account identifies one place where harm can continue operating after external causes have been reconstructed inside you. It does not relocate blame merely because the later operation is internal.

The claim is not:

> you suffer because you think wrongly.

The claim is that suffering can become epistemically structured, and that the structure may have causes far outside the individual.

5. Corrigibility as the Moral Foil

Once closure is the mechanism, its positive contrast becomes clearer.

Human flourishing emerges where the epistemic cycle remains open. Problems can be formed. Tentative solutions can be created. Criticism can enter. Explanations can change. Successor problems can emerge.

A flourishing life is not a solved life. It may contain more problems than a stagnant one.

The difference is that the problems can move. Suffering often becomes especially intense when they cannot. No acceptable conjecture appears. Exit is blocked. The criterion cannot be revised. The problem is imposed. The explanation cannot be criticized. Nothing changes the landscape.

You are not merely in pain. You are trapped inside a problem that has become structurally closed.

The structural contrast may be abstracted as:

$$\text{flourishing} \Rightarrow \text{preserved epistemic movement}$$

while

$$\text{evil-preserving closure} \Rightarrow \text{blocked correction that preserves avoidable suffering.}$$

This is not literal. But the arrows highlight a contrast: whether a harmful problem can still move epistemically.

It also changes what peace can mean.

Peace need not mean emptiness, the absence of conflict, or the absence of difficult problems. It may be what becomes possible when conflict can transform: when an explanation is not final, when criticism can enter without destroying you, when a self-model can change, when a problem can generate a successor, when incompatible structures do not have to remain in permanent war.

Some suffering has biological or material causes that epistemic openness cannot dissolve. But the structural difference is recognizable.

You already know the difference between:

> I do not know the answer yet.

and:

> There is no way out.

The first can be difficult. The second can become despair.

6. The Self-Regulation Objection

The positive contrast does not remove the darker objection. It sharpens it.

If you can become trapped in rumination, destructive self-models, conflicting explanations, and compulsive interests, why should anyone trust you with freedom? If one mind cannot regulate itself, how can billions of such minds coordinate?

The epistemic account changes the unit of analysis. The question becomes: what environment of objective knowledge is the present individual attempting to regulate themselves inside?

You are not an isolated biological processor. You live inside inherited explanations, status hierarchies, moral narratives, economic structures, psychological categories, media systems, family expectations, institutional roles, and technologies. Many internal conflicts may be reconstructions of social objective knowledge.

Your suffering is real. But it does not follow that your freedom caused it. Constraint can even be misread as evidence for more constraint.

Abstract it:

$$\text{epistemic constraint} \rightarrow \text{persistent unresolved problem} \rightarrow \text{suffering.}$$

An observer sees the suffering and concludes:

$$\text{suffering} \rightarrow \text{person requires greater regulation.}$$

The result is more constraint. The mechanism producing the problem is intensified.

We have already encountered versions of this: the soul-crushing job, the child whose curiosity collides with institutional sequencing, the lonely person trapped inside a bad problem formulation, the dissenter whose criticism is treated as pathology, the person who cannot stop ruminating because every available explanation returns to the same closed problem.

The surface state can be read as evidence of defective human nature. It may also be evidence of what happens when the epistemic cycle is obstructed.

That should make us cautious about claims concerning human nature. You do not encounter your own capacities in a vacuum. You live inside an enormous inherited environment of objective knowledge: centuries of authority, moral cosmologies, bureaucratic categories, market structures, education systems, media incentives, psychological theories, technologies of attention, legal systems, and family norms.

Any observation of human nature therefore begins with people already embedded in epistemic structures. Some are liberating. Some constraining. Some contradictory. Some obsolete.

So when you find yourself anxious, conformist, punitive, passive, or self-destructive, the epistemic question is not only what trait or tendency you possess. It is also:

> Which parts of this state arise from your present problem situation, and which are inherited objective knowledge reconstructed inside you?

The distinction is difficult. That does not make it optional.

7. Objective Knowledge Carries Harm—and Correction

A harmful explanation does not remain private once expressed. It can become doctrine. Policy. Tradition. Curriculum. Diagnosis. Slogan. Algorithm. Procedure. Meme. It becomes objective knowledge.

Another mind can then reconstruct it. Abstractly:

$$M_1 \to O_E \to M_2 \to O'_E.$$

No collective mind is required. What looks like social evil can be understood as many solitary minds reconstructing, preserving, modifying, and enforcing harmful explanatory structures. This gives objective knowledge a dark side. It can preserve error beyond the life of the person who created it.

But the same is true of criticism. A method can be written down. A right can be institutionalized. A procedure can be redesigned. Testimony can become evidence. A scientific correction can enter textbooks. A moral insight can become a norm. Our institutions of criticism.

You do not need to rediscover every error from scratch. Objective knowledge allows corrections to travel. That is why preserving error correction matters morally. The same infrastructure that scales harmful explanations can scale their criticism.

The relevant dynamics therefore do not divide neatly into pure minds and evil institutions. Harmful explanations and their criticisms can both become objective, travel between people, and be reconstructed inside fallible persons.

That changes the self-regulation objection. You have never regulated your epistemic life unaided. Language helped. Writing helped. Books helped. Friends helped. Science helped. Institutions helped. Therapy helped. Technology helped.

The relevant unit has always been something like:

$$\text{person} + \text{objective knowledge} + \text{tools}.$$

Tools are one of the interfaces through which objective knowledge becomes usable in your problem situation. The question is therefore not whether a solitary biological processor is sufficient for moral life.

It never was.

The question is which forms of objective knowledge, embodied in which tools and practices, enlarge your capacity to expose harmful explanatory structures without simply replacing one authority with another.

That question becomes newly concrete because the objective environment around you has recently and profoundly changed.

8. A Responsive Surface for the Solitary Mind

For most of history, making an internal problem explicit required simple interfaces. Conversation. Writing. Teaching. Friendship. Books. Each could help you externalize part of an epistemic landscape.

Each also had boundaries. A book could not reorganize itself around your current problem. A library could not ask what changed in your formulation. A therapist had finite time and a professional frame. A friend carried their own problem situation. An expert answered from within a domain.

Now you increasingly have access to a strange new surface. The distinction must remain clear: conversational AI is not another person merely because it responds in language. Fluency does not establish a subjective life. AI can retrieve, compare, reorganize, restate, synthesize, and execute methodologies. None of that alone establishes an artificial solitary knower.

But a tool does not need to be a person to change your epistemic environment. Writing never became conscious. Libraries never became persons. Calculators never understood arithmetic. Objective knowledge can transform cognition without becoming a knower.

AI belongs in that lineage. Its difference is dynamic responsiveness. You can bring an inchoate problem into conversation. Something feels wrong. A relationship makes no sense. A job feels intolerable. A self-model keeps collapsing. Two explanations conflict. An interest has become consuming. A moral judgment feels too simple.

You may not yet know the problem. Conversation externalizes fragments. The tool can preserve them, compare them, restate them, expose contradictions, offer alternative formulations, and reorganize objective knowledge from distant domains around the problem you are trying to articulate. You encounter parts of your own epistemic structure in a new form.

Abstract it:

inexplicit tension → externalized language → articulated problem → criticism → possible revision.

The AI has not solved the moral problem. It has provided a surface on which you can inspect it.

This also helps explain one striking use of conversational AI: deeply personal conversation. People bring it relationships, anxiety, work, identity, loneliness, grief, decisions,

self-conceptions, and unresolved conflicts. That should not be confused with professional therapy. The epistemic significance is different.

The relevant fact for this argument is that people can bring the technology intimate problem situations that may otherwise be difficult to externalize. One reason may be the breadth of the interface. A conversation can move from work to childhood to philosophy to music to morality without needing to fit a professional category.

Your problem world is not organized like an academic department. You are the common denominator.

The important distinction is therefore not whether the tool can supply the right problem for you. That would recreate a false authority.

The more interesting use is almost the opposite. You arrive with the problem, even if only implicitly. The tool helps make it explicit. You reject a formulation. The problem changes. Objective knowledge is reorganized around the new state. You criticize again. The process continues.

That is structurally different from a system that says:

> Here is the answer you should accept.

The epistemic opportunity lies in saying:

> Here is a form in which your problem can become criticizable.

This inquiry did not begin with a complete multi-part architecture. One problem created another. Behavioral genetics created a problem about psychological explanation. Psychology created a problem about individuality. Individuality created problems about society, freedom, abstraction, consilience, criticism, problem formation, coordination, and now evil.

The criteria changed. Some formulations were rejected. Some concepts became more precise. Some problems were abandoned. Others deepened.

The AI did not supply the governing path. The evolving human problem situation did. The technology allowed objective knowledge to be repeatedly reorganized around that path as it changed.

That is not proof of the theory. There's no such justification. Rather, it is an instantiation consistent with the thing being explained; epistemology in motion.

The symmetry still matters. AI can reinforce bad explanations. It can flatter. It can personalize ideology. It can deepen closure. Institutions can use it to make people more legible and controllable. A responsive epistemic surface can therefore serve correction or become another mechanism of authority.

But the possibility space has changed. You can now access a responsive tool capable of reorganizing large bodies of objective knowledge around an idiosyncratic problem situation.

For this inquiry, that change matters because it alters the conditions under which a harmful explanation can become explicit and criticizable.

9. The Epistemic Problem of Evil

The ancient explanation said, in effect:

> Evil lives inside the person, therefore the person requires external control.

The epistemic account is both darker and more optimistic.

Yes, the possibility of evil is internal to us. The capacity to create, inherit, and stabilize explanations is unbounded. You can become closed. You can caricature another person. You can rationalize. You can internalize an authority. You can become trapped in a bad problem.

But the same explanatory capacity also makes criticism and revision possible.

You can discover the caricature. The self-model can become a problem. The inherited authority can be criticized. The bad problem can be reformulated. The interest can change. The explanation can fail. A successor problem can open.

That is the foil. Not innocence. Corrigibility.

The problem of evil can therefore be stated differently:

> Why does avoidable suffering persist in a world containing people capable of explanation?

The tentative answer is not that humans contain a permanent evil substance. It is that fallible explainers can create, inherit, and stabilize harmful knowledge.

The causal sequence is now visible. Harmful explanatory content can be reconstructed in you. It can become insulated from criticism. Power or internal self-reinforcement can stabilize it. Objective knowledge can preserve and propagate it. The resulting closure can preserve avoidable suffering.

That is the explanatory danger. But the same epistemology explains why evil need not be final. Knowledge remains revisable. Criticism can travel. Objective knowledge can carry correction. Tools can enlarge your capacity to inspect the structures you inherited. Freedom can preserve the problem space in which moral error becomes visible. We can build lasting institutions of criticism.

The moral imperative that follows is not:

> obey the correct authority.

It is:

> preserve the means of error correction.

That principle does not solve morality. It can't tell you what to do. But it may preserve the conditions under which a moral error can become visible as an error.

Now return to you. You have probably classified someone too quickly. You have probably carried a self-explanation that later collapsed. You have probably inherited a moral rule without knowing where it came from. You have probably experienced the difference between a problem that could move and one that felt closed.

You may have felt relief when a sentence, a person, a book, or a tool made a hidden tension explicit. You may have caused harm while solving the wrong problem. You may have been harmed by someone else's explanation of who you were.

None of this is evil by essence. There's no such thing.

It makes you a fallible explainer. The epistemic possibility of evil travels with it. So does the possibility of moral growth.

The older fear was that human nature must be restrained because evil lives within it. The more demanding possibility is that what lives within us is not fixed evil but a universal capacity for explanation whose products can become harmful, closed, and enduring.

And if you are reading this while interrogating the argument through your own questions, perhaps with an AI beside you, the process is not merely something being described.

The harmful explanation can become visible. The problem can move. The next explanation can be different. So can you.

That leaves the successor problem.

If objective knowledge can preserve corrigibility or destroy it, and if increasingly responsive tools can participate in either process, what distinguishes an epistemic environment that enlarges your capacity for error correction from one that merely becomes a more adaptive form of authority?

XII. Our Justified World



The epistemology explored here thoroughly rejects justification. Knowledge does not become secure through the accumulation of supporting observations. Agreement, confidence, probability, consensus, and coherence are not foundations of truth. Knowledge evolves through problems, conjecture, criticism, error elimination, and the continued survival of explanations that have not yet been eliminated.

Yet ordinary life seems to pull in the opposite direction. We ask for reasons, examples, evidence, confidence, likelihood, and agreement. We teach students to defend theses. We build institutions that reward supported conclusions and statistical tools that assign numerical confidence. We built artificial intelligence systems inside a culture saturated with these justificatory forms.

And when an explanation fits many things we already know, something happens inside you:

> it makes sense.

That feeling is not trivial. It can mark successful reconstruction. It can mark explanatory integration. It can mean that things which previously seemed disconnected have become intelligible within one structure.

It can also be mistaken for confirmation.

So the problem is not merely whether justification fails logically. The problem is why it remains so psychologically and culturally natural:

> If knowledge does not grow by justification, why does justification feel so natural?

Calling people irrational would explain nothing. It would take the pattern to be explained and turn it into a primitive of human nature.

A better route is to separate two things that usually arrive together in experience. One is the logical structure by which explanations survive criticism. The other is the biological, cultural, and technological machinery through which you actually think. The first can be asymmetric even when the second does not make the asymmetry feel natural.

You can understand general relativity without intuiting curved spacetime. You can read even though your visual system did not evolve for printed language. In the same way, you can understand that matching observations do not justify an explanation. Your mind, culture, and tools can still repeatedly produce the feeling that they do.

An orientation toward disconfirmation may therefore be an achievement. Not because criticism is merely psychologically unpleasant, but because the objective knowledge required to practice it had to be created.

1. How Justification Became Common Sense

The demand for justification is much older than formal epistemology. Long before philosophers asked how knowledge should be grounded, communities already possessed practical structures that authorized belief. Truth could be revealed. A sacred text could disclose it. A priesthood could interpret it. A ruler could embody it. Tradition could preserve it.

Abstract it.

The relation was straightforward:

$$\text{truth} \rightarrow \text{authority} \rightarrow \text{belief}.$$

Authority did not merely communicate an explanation. It supplied entitlement to believe.

This matters because justification is not only a technical problem in philosophy. It is part of the architecture by which objective knowledge has historically moved through institutions and into individual minds.

Observation changed that architecture. Instead of asking only:

> Who says this is true?

people increasingly asked:

What do we observe?

That change weakened inherited authority and created new forms of criticism. But the demand for epistemic entitlement did not necessarily disappear. It could move from the source of the claim to the observation alleged to support it.

The new question became:

How does observation justify belief?

The authority of revelation could become the authority of observation. The source changed. The justificatory function remained.

The familiar idea that knowledge is justified true belief captures why the function is attractive. You do not want merely to possess a belief. You want it to be true, and you want some relation between the belief and the world that licenses treating it as more than a conjecture.

But finite observations do not entail universal explanatory truth. No number of matching observations can deductively generate the theory that explains them. That is not a practical inconvenience. The entailment is absent.

The anti-justificationist response is therefore not:

certainty would be ideal, but unfortunately we cannot have it.

It is more radical:

justification was never required for knowledge growth in the first place.

You encounter a problem. You create an explanation. You criticize it. You act on it. You replace it when criticism exposes a better problem situation. Nothing in that sequence requires a state called justified belief.

Certainty, however, is not the only form justification can take. Once certainty became difficult to defend, the same function could reappear in weaker language:

- probably true;
- highly likely;
- very high confidence;
- the weight of the evidence;
- backed by consensus.

The vocabulary changes from certainty to degrees, while the practical conclusion can remain:

I now possess enough support to treat this explanation as secure.

Probability enters here, but it creates another problem.

Probability is a powerful formal tool. It has uses in science, engineering, finance, communication, logistics, risk management, and many other domains. But formal probability is objective knowledge. It is a cultural invention, recent relative to biological evolution. Whatever biological machinery enables you to create explanations did not evolve by performing probability calculus.

Once learned, however, objective knowledge can become part of the machinery through which you think. If you are trained probabilistically, questions about likelihood, confidence, expected value, and distributions can become cognitively automatic. The tool stops presenting itself as a tool.

Then a dangerous compression becomes possible:

> uncertainty is probabilistic.

That sentence sounds more general than the formal structure permits. Probability assigns weight within a represented possibility space. It presupposes outcomes, variables, states, models, or some other structured set over which the distribution is defined.

Show me where that matters.

Take a deck of cards. The deck is finite. The rules are specified. The relevant possibilities are sufficiently closed for calculation.

Now take a scientific anomaly. The theory may not yet exist. The relevant variables may be wrong. The categories may be wrong. Even the problem may be wrong.

The difference is not between precise and imprecise reasoning. It is between assigning weight inside a represented space and creating knowledge that can change or dissolve the represented space itself.

2. When the Problem Changes

The distinction becomes clearest when the evaluative criteria move.

Explain.

Suppose your problem is:

> Which job should I take?

You may initially compare salary, prestige, and security. If those criteria stay fixed, the problem can be represented as a choice among known options under a known ranking structure.

Then a different question appears:

> Why am I optimizing for prestige?

The available jobs have not changed. What changes is the problem. Autonomy, interest, time, creativity, or relationships may become relevant because the criterion by which the options were being ranked has itself become criticizable.

You have not merely selected a different item from a fixed menu. You have created a different decision problem. This matters because open-ended explanation does not only add possibilities. It can alter what counts as a possibility, what counts as a relevant variable, and what counts as success.

That changes the relation between knowledge and action.

A common defense of justification is practical:

> Without confidence, how could you act?

Abstract it.

The question assumes a particular sequence:

$$\text{assign credence} \rightarrow \text{cross confidence threshold} \rightarrow \text{act}.$$

But action can be organized by a different epistemic structure:

$$P \rightarrow TS \rightarrow A \rightarrow EE \rightarrow P_2.$$

You have a problem, P. You conjecture a tentative solution, TS. You act, A. The consequences expose errors, EE. The problem changes into P_2.

This sequence does not require final epistemic security. Action is one way a conjecture enters reality strongly enough to encounter criticism.

Newtonian mechanics makes the point concrete. It is not a true account of physical reality, yet it remains useful for enormous classes of problems. Within regimes where relativistic corrections are negligible, Newtonian models can solve the narrower problem well enough to support action.

That does not imply instrumentalism.

What?

The view that ideas and scientific theories are only tools for prediction. Reality still explains why the approximation succeeds where it does. The narrower claim is that truth and local problem adequacy are not identical.

An explanation can therefore be false in general and still be the right problem-solving instrument locally.

Explain.

Practical action does not require the conclusion:

> this theory is finally justified.

What matters instead is more specific:

> What problem does this model solve, where does it fail, and do those failures matter here?

That question forces a distinction the language of "evidence for a theory" often blurs. Separate the abstractions.

An **explanation** asserts dependencies in reality. It says how and why something happens because the structure of the world has a particular form.

A **model** represents some part of that structure for a particular problem. It may simplify, idealize, ignore variables, or hold others fixed.

An **experiment** creates a test situation.

An **observation** is what is encountered within that situation.

Compress it.

Their relation can be represented as:

$$E \to M \to X \to O,$$

where E is the explanation, M the model, X the experiment or test, and O the observation.

Criticism travels back through this chain cautiously:

$$\underbrace{\text{criticism of O} \to \text{criticism of X} \to \text{criticism of M} \to \text{ criticism of E}}_{\text{possibilities}}$$

The word *possibilities* is load-bearing. An unexpected observation may reveal a bad measurement, a bad model, a false auxiliary assumption, a regime where an approximation fails, or a genuine explanatory failure. The observation does not announce which part of the structure broke. Everything is conjectural.

Two rival explanations can also produce nearly identical models over a restricted domain. Both may fit the same observations. If so, those observations do not discriminate between the explanations.

So:

$$\text{model success} \neq \text{explanatory truth}.$$

This isn't just about science. The distinction becomes clearer when the example of everyday examples enters the argument.

3. Why Examples Feel Stronger Than They Are

Suppose I am trying to explain a novel abstraction to you:

> epistemology changes how we explain people.

The sentence is too compressed. You may not directly possess the explanatory structure I have in mind. So you say:

"For example?"

Now I give you the personality-trait case, the lonely young man, the child in school. Each example removes some possible reconstructions of what I meant. The examples help you rebuild enough of the structure inside your own mind for the abstraction to become intelligible and criticizable.

That operation is useful enough to name. When one mind uses language, cases, and distinctions to let another mind conjecture an explanatory structure, the second mind is performing **reconstruction**.

The order matters. The examples do not mechanically generate the explanation. They constrain your attempt to reconstruct it.

Abstract it.

The logical relation:

$$O_1,\ O_2,\ O_3$$

does not entail:

$$E.$$

The examples help you reconstruct E. They do not justify or confirm E.

So examples are epistemically irrelevant?

No. That would erase the very function we just identified. Examples can make an explanation communicable. They can reveal what a compressed abstraction means in practice. They can also participate in tests once an explanatory relation has been specified. What they cannot do merely by fitting is derive the explanation that makes them intelligible.

The same distinction becomes visible when an example is missing. Someone offers a high-reach explanation. You ask for an instance. The speaker struggles.

The failure may matter communicatively. It may diminish the rhetorical impact. Perhaps the abstraction is too vague to reconstruct. Perhaps the speaker has not connected it to any usable model. Those are real criticisms. But inability to improvise an example under conversational pressure does not by itself eliminate the explanation. Example formulation and explanatory content are different capacities.

Now the psychological problem sharpens.

Once enough examples fit, you begin to feel:

> yes, I understand.

Then, easily:

> yes, that is right.

Those are different achievements.

The first can mark successful reconstruction. The second is an epistemic judgment about the explanation.

The transition is easy to miss because the same mechanism that lets you inhabit another person's abstraction also produces a feeling of fit. A meeting of the minds. The better the reconstruction works, the more coherent the surrounding observations can appear.

Keep the relation and drop the incidental detail:

$$\text{successful reconstruction} \neq \text{successful error elimination}.$$

The feeling of understanding is not meaningless. It may show that explanatory structure has actually been transmitted. But transmission and truth are different relations.

Observations make the transition easier because their descriptions are comparatively plastic. The same event can be described as:

> The raven is black.

or:

> The raven is not non-black.

The event has not changed. The linguistic surface has.

A more consequential case has the same structure. One feature of an intervention can be emphasized, another omitted. The same event can be embedded in rival narratives because the observation does not contain its own explanatory role.

Explanations are rigid in a different sense. They assert dependencies. If an explanation says:

$$A \to B,$$

then neither

$$B \to A$$

nor

$$\neg A \to \neg B$$

follows.

Explain.

The direction of dependence is part of the explanation. Rival explanations may accommodate the same observations while making incompatible claims about what depends on what. Their conflict appears in the structure that the observations alone leave underdetermined.

This suggests a neat procedure: assemble rival explanations, derive their consequences, and test where they diverge.

No. That begins too late.

Often there is enormous data, observation, correlation, and measurement while there is still no good explanation. The difficult part of knowledge creation is frequently creating even one conjecture with enough structure to criticize. It may be exceedingly difficult to formulate even one good explanation for a given problem situation.

Two serious explanatory rivals are often an epistemic luxury. Their incompatible dependencies may make sharp tests possible, but criticism does not require a menu.

If you have only one serious conjecture, criticism can still reach into its structure. A dependency may contradict another part of the explanation. An assumption may be doing work the explanation has not acknowledged. A predicted observation may fail to occur. The explanation may solve one problem only by losing reach somewhere else. The general requirement is not multiple options. It is vulnerability to error.

This is already enough to explain one part of the felt asymmetry. Confirmation can arrive through recognition: another fitting case, another successful reconstruction, another observation that can be placed inside the model. Criticism requires exposing the structure strongly enough for something in it to fail or become problematic.

4. Why Confirmation Is Cheap

The distinction between explanation and model offers a more precise account of memetic compression.

A complex explanation may contain many dependencies, qualifications, boundary conditions, unresolved criticisms, and rival interpretations. Carrying all of that explicitly is cognitively expensive. So objective knowledge is often compressed into something smaller: a phrase, a slogan, a category, a story, a memorable case.

Take:

> people need strong leaders.

The sentence can stand in for a much larger explanatory structure about coordination, uncertainty, hierarchy, authority, competence, and social order. The compressed form is

easier to remember, easier to transmit, and easier to instantiate. But it contains much less structure.

That tradeoff is unavoidable. Compression is not itself an error. You cannot keep every dependency of every explanation available to attention at once. The vulnerability appears when the compressed object stops functioning as an index into a larger explanation and begins functioning as the explanation itself.

Once that happens, confirming observations are readily captured.

The meme:

> kids today are spoiled

finds instances.

The meme:

> experts know best

finds instances.

The meme:

> authorities cannot be trusted

finds instances.

The examples differ, but the mechanism is the same. The available compression organizes salience. Events that fit it are easy to classify and remember. The selected observations then return to consciousness already shaped by the model that made them salient.

Abstract it:

$$M \rightarrow \text{salience} \rightarrow O_{\text{selected}} \rightarrow \text{felt confirmation.}$$

No formal induction is required. You only need to keep recognizing instances of an already available compression.

And then you show someone else. The mechanism does not stop at the boundary of one mind.

So if the meme is false:

> Bullshit travels faster than the truth.

You just recalled a meme that makes a meme out of the Bullshit Asymmetry Principle.

Brandolini's law is not an epistemic law. It is a set of weakly tied conjectures. It asserts that falsehoods are usually simple and truth complex. That people prefer fast certainty over ambiguity. That media reward falsehood over accuracy, and that we can become overwhelmed in a glut of false claims.

But it does describe an asymmetrical epistemic environment. It is the recursive application of confirmation operating within our justified world. A single disconfirming observation may stop bullshit in its tracks.

I see what you did there.

Much ordinary communication consists of trying to make another person see what you have come to see:

> You won't believe what I saw.
>
> Did you see that?
>
> You have to hear this.

An observation, story, or compressed explanation changes your epistemic landscape. You then try to reproduce some of that change in another mind. We do not need a biological theory of the compulsion in order to notice the structure. People persistently try to render actions, beliefs, and events intelligible to one another.

Someone behaves unexpectedly. You ask why. A friend makes an odd decision. You reconstruct the problem that might make the action intelligible. You do the same to yourself:

> Why did I do that?
>
> Why does this matter to me?
>
> What am I actually worried about?

This activity is older and broader than the philosophical concept of justification. It is explanation-seeking as social reconstruction.

But explanation can slide into justification because intelligibility often feels like validation. When an action becomes understandable, the subjective change can be mistaken for a change in its epistemic or moral status. When a belief becomes reconstructable, the same thing can happen.

Now add agreement.

Suppose I tell you what I saw and you reply:

> Exactly.

Two solitary minds have reconstructed compatible versions of the same abstraction. That is genuine social consilience: successful transmission between solitary minds.

But transmission is still not truth. Agreement may indicate shared background knowledge, shared compression, successful reconstruction, or shared error. Yet there

is a large subjective difference between nobody seeing what you mean and everyone seeing it.

That difference gives justification a social charge. A successful compression can also become social currency. If a phrase, example, or story is memorable and easy to instantiate, recognition by others can bring status. The sender may appear perceptive, funny, morally alert, correct, or simply inside the group.

Truth-seeking and status-seeking do not need to be consciously opposed for this to matter. You can genuinely want others to see what you saw while also receiving status from successful transmission.

Social media changes the scale and speed of the mechanism. Recognition becomes visible, immediate, public, and countable through likes, shares, replies, followers, and reposts. Those signals are not epistemic validation. But they can become adjacent to the same subjective machinery.

Abstract it:

$$\text{compression} \rightarrow \text{propagation} \rightarrow \text{recognition} \rightarrow \text{status} \rightarrow \text{more propagation}.$$

A vivid example can therefore travel faster than the explanation that would determine what the example actually bears on. Public discourse becomes saturated with single instances: one video, one scandal, one image, one anecdote, one outrageous quotation. Each can be used as a lightweight model for a much larger class, and the relationship between observation and explanation is not 1:1.

The informational environment changes the costs. Compression is cheap. Propagation is cheap. Recognition is rewarded. Reconstructing a large explanatory surface and exposing it to criticism remain comparatively expensive. The result can be an environment rich in apparent confirmation and poor in explanatory criticism.

No social mind is required. The epistemic act still occurs inside solitary minds. The network changes frequency, salience, cost, and reward. Each individual still reconstructs the compression for themselves.

This gives us a biological, cultural, and technological route by which confirmation can become easy to feel.

It still leaves one question untouched:

> Why does criticism have any special epistemic force at all?

5. The Asymmetry

The answer is logical rather than psychological.

Suppose an explanation entails an observable consequence:

$$E \rightarrow O.$$

If the test yields:

$$\neg O,$$

then, assuming the relevant premises of the test are retained:

$$\neg E$$

follows by modus tollens.

What?

The inference is deductive. If the premises are true, the conclusion cannot be false.

Reverse the result.

Suppose the explanation entails O, and you observe O:

$$E \rightarrow O$$

and

$$O.$$

Those premises do not entail E. Other explanations may also entail O. Inferring E would affirm the consequent.

That is the asymmetry. It is not a preference for negative thinking. It is not a personality trait. It is not a cultural taste for criticism.

But experiments often succeed. That matters. The error is in saying what the success does.

If an explanation survives a severe attempted falsification, we may call it **corroborated**. The term records that the explanation survived criticism that could have exposed a problem. It does not mean the matching observation derived the explanation, added truth to it, or justified it.

Corroboration is survival of criticism, not accumulated support. This leaves empirical rigor intact. Rejecting justification does not require abandoning experiment, measurement, observation, prediction, statistics, model comparison, or replication. Those practices can expose discrepancies, characterize restricted problems, and eliminate false models.

What changes is their epistemic interpretation. A measurement can reveal a failure. A statistical method can describe a represented problem. An observation can participate in a test. None supplies a deductive route from finite instances to universal explanatory truth. The tools remain. The justificatory interpretation is what disappears.

There is also a positive constraint that does not reduce to falsification.

An explanation about one part of reality exists in the same world as explanations about everything else. If it solves one problem only by contradicting powerful explanations

elsewhere, the conflict itself becomes a problem. Compatibility across explanatory domains will be called **consilience**.

Consilience and falsification therefore do different work. Falsification asks where an explanatory structure breaks under a test. Consilience asks whether the structure can coexist with other powerful explanations of the same reality without hidden contradiction.

Neither justifies the explanation. Both constrain it.

The distinction matters reflexively because epistemology has universal reach. The same explanation has been carried through behavioral genetics, psychology, society, institutions, freedom, abstraction, consilience, criticism, problem creation, coordination, evil, and now justification.

That increasing coherence can feel compelling. It may be genuine consilience. It may also generate exactly the misleading feeling this essay is trying to criticize.

So the relevant test is not:

> How many more examples fit?

It is:

> Where does the explanation break?

That question returns us to probability, because a tool can become hardest to criticize precisely when it disappears into ordinary thought.

6. When a Tool Disappears into Thought

Formal probability is objective knowledge. Once integrated into cognition, however, it can stop appearing as an external technique and begin structuring the questions that seem natural.

You may find yourself asking automatically:

- how likely is this?
- what is the confidence level?
- what probability should I assign?
- which option has the highest expected value?

Within represented spaces, those questions can be exactly the right ones.

The danger appears when the question itself is open. A probabilistic frame can then import assumptions that have not earned their place. The relevant options are treated as known, the variables as stable, the evaluative criterion as fixed, and the hypothesis space as already containing what matters.

The tool becomes invisible. The problem is then reformulated until it fits the tool. Modern data environments intensify that pressure. They produce measurements, metrics, behavioral traces, sensor data, correlations, and forecasts in volumes that earlier cultures could not produce.

Observation can therefore become abundant while explanation remains scarce. A field can measure more and more without having a better answer to:

> What is actually going on?

That distinction creates a technological waypoint.

Inductive and probabilistic methods may offer predictions. But if explanatory knowledge is not logically derivable from prior observations, then predictive success alone does not explain the universal explanatory capability attributed to persons.

If AGI is to possess that epistemic capability, the problem is not merely:

> how does the system predict well?

A further problem appears:

> how does the system create explanatory knowledge that is not logically derivable from the observations already represented?

Nothing here settles competing AI programs. The distinction is narrower. Inductive competence and explanatory creativity are not the same explanatory object.

Now the technological problem folds back into the cultural one. Current AI systems were created inside societies saturated with support, evidence, confidence, argument, defense, consensus, probability, and optimization. Their training material contains essays defending theses, legal arguments, political rhetoric, marketing, academic papers, textbooks, apologetics, social media, and opinion.

Much recorded language is justificatory in form. So AI inherits more than objective knowledge. It also inherits cultural forms for defending what is already being said.

Turn that into a mechanism.

A user supplies a compression: a belief, thesis, suspicion, ideology, diagnosis, or preferred explanation. The system can search the available objective knowledge for material that can be attached to it: examples, arguments, analogies, citations, historical parallels, theoretical vocabulary, reconciliations of apparent contradictions. The output becomes richer. The original compression remains unchanged. The user experiences increasing confirmation.

AI becomes a justification machine. AI automates justificationism.

7. AI in Both Directions

The asymmetry is economic as well as logical. You can ask a machine for twenty supporting arguments, twenty examples, ten sources, several analogies, a stronger rhetorical version, and a more sophisticated theoretical defense in very little time. The cost of generating reasons collapses.

But genuine criticism still requires something more specific than producing the opposite list. The explanatory structure has to be exposed. A dependency has to fail. An assumption has to be located. A rival explanation has to make a different dependency visible. A model boundary has to matter to the problem.

A system that makes support abundant can therefore amplify confirmation without changing the logic of confirmation at all. That is one direction.

Now reverse it.

AI is treated as a technology of retrieval, synthesis, compression, and reorganization. The present problem reveals another possible function.

A large explanation exceeds what you can keep available in working attention. It may contain many dependencies, qualifications, rival accounts, historical revisions, unresolved criticisms, and model boundaries. Unaided, you compress.

Until recently the practical choice was often between carrying a small compression biologically and relying on an external authority, text, or institution to carry more of the structure.

AI creates another possibility: externalize a larger explanatory surface and recover the relevant parts when the problem calls for them.

This is more than memory of propositions. What matters is preserving relations. Which claims depend on which others? Which assumptions are local? Where does a model stop working? Which criticism remains unresolved? Which rival explanation covers the same observation differently?

An external system that preserves those relations can function as **explanatory memory**. That term names a specific capability, a specific epistemic medium. The machine stores or reconstructs enough of the explanatory surface that a local compression can later be reopened for criticism.

Compression therefore remains necessary but need not be irreversible:

$$E \to M_c$$

for local cognitive use, while a route back remains available:

$$M_c \leftrightarrow E.$$

The bi-directionality changes what compression can do. A meme can lead back to the larger explanation. An example can lead back to the dependency it instantiates. A slogan can be expanded into claims, assumptions, boundaries, and rival accounts.

Compress and decompress.

That operation matters especially for large inherited abstract structures: religions, political ideologies, moral systems, economic theories, legal traditions. Few individuals can reconstruct such systems at full resolution. Historically, authorities, canonical texts, doctrines, summaries, slogans, and interpreters carried parts of the explanatory burden. That solved a real cognitive-cost problem. Someone or something outside you carried structure you could not keep active yourself.

It also created a route to closure: the verdict could arrive before the structure that made the verdict criticizable. AI can lower the cost of asking for that structure back.

You can ask which claims depend on which others, which parts are empirical, which are moral, which are historical, where rival explanations cover the same observations, and where those explanations actually diverge.

This is **decompression**: not recovery of some perfect original, but expansion of a compressed representation into a richer explanatory structure with more explicit points of possible criticism.

The relation to authority changes because you can increasingly request:

> Show me the explanation, not merely the verdict.

This does not make expertise irrelevant. It changes how much explanatory surface can be made available to the singular person confronting a problem. To you.

Social media and AI therefore create opposite transformations around the same object. Social media made compression and propagation extraordinarily cheap:

$$\text{explanation} \rightarrow \text{meme} \rightarrow \text{mass propagation}.$$

AI can sometimes make the reverse traversal cheaper:

$$\text{meme} \rightarrow \text{expanded model} \rightarrow \text{underlying explanations} \rightarrow \text{criticism}.$$

The symmetry is imperfect. AI can also generate memes, propaganda, and confirmation loops. The capability does not determine its use.

But a new operation is available. You can take a viral example and ask what explanation would make the observation relevant. Then the structure can be reopened: what rival explanation accommodates it equally well, which assumptions have been compressed away, and does the example actually discriminate between explanations?

AI can also reduce the cost of holding rivals open. You often settle on one working explanation because maintaining several large alternatives is expensive. An external system can preserve explanation A, explanation B, explanation C, together with their assumptions, incompatibilities, domains of success, and shared observations.

The resulting question changes from:

> Which one currently feels most coherent?

to:

> Where do they diverge?

That can extend your epistemic reach. It does not settle the status of the machine.

AI can carry objective knowledge, reorganize explanatory structure, generate candidate criticisms, preserve alternatives, and reopen compressed models. None of those capabilities alone establishes that the machine is an AGI.

The epistemic subject in this inquiry remains you: the individual for whom something is a problem, who can reject the framing, evaluate criticism, and create the next problem. The machine can enlarge the explanatory surface available to you. It does not thereby become the solitary knower whose problem the surface serves.

To think otherwise is to misunderstand what our confirming observations can support, without an explanation for how.

8. You Are Inside the Asymmetry

The same technology can therefore participate in opposite epistemic processes.

As justification:

$$\text{AI} \rightarrow \text{more support} \rightarrow \text{closure.}$$

As explanatory memory:

$$\text{AI} \rightarrow \text{more preserved structure} \rightarrow \text{more opportunities for criticism.}$$

The technology does not determine which process occurs. The difference appears in the problem you bring to it and in what happens to the explanatory structure afterward.

If the initial belief remains fixed, the system can be asked to find support, reconcile objections, and generate confirming cases. The explanatory surface expands while the underlying compression remains protected.

The operation can be reversed. The system can expose dependencies, recover model boundaries, preserve rivals, and identify where an explanation could fail. The same increase in available structure now creates more points of criticism.

The distinction is not between optimistic and pessimistic uses of AI. It is between operations that increase closure and operations that increase vulnerability to error.

Return to you. Why does justification feel natural? Several mechanisms now converge.

Objective knowledge carries justificatory traditions. Formal probability supplies powerful reasoning inside represented spaces and can become so cognitively familiar

that its boundary conditions disappear from view. Examples help solitary minds reconstruct one another's abstractions. Successful reconstruction produces a real feeling of fit. Compression makes large explanatory structures cheap enough to remember and transmit. Once a compression is available, salience supplies further instances. Social recognition rewards successful transmission. Social media quantifies that reward. AI can make supporting reasons abundant almost without limit.

None of these mechanisms creates the missing logical entailment from matching observation to explanatory truth. Together they may explain why the absence of that entailment need not be psychologically obvious.

This also explains why criticism can require effort without treating that effort as evidence of a defect in human nature.

The machinery doing ordinary cognitive work solves other problems. Reconstruction lets you understand another mind. Compression reduces cognitive cost. Stable models make action possible. Shared abstractions permit coordination. Recognition helps information propagate. Probabilistic tools solve real closed-space problems.

Criticism can work against the local success of those mechanisms. It can require you to reopen a compression that is already useful, expose an assumption that stabilizes action, or preserve a rival that makes coherence harder to maintain. It can also disturb a social identity built around the compression, destabilize social membership organized around it, reopen a problem that had seemed settled, or destroy a simplification that was doing useful work.

The cost is therefore intelligible. Criticism is a cultural technology for forcing explanatory structures back into contact with possible error.

Now turn the criticism back on this essay.

By this point, the epistemology itself may feel increasingly natural. The examples fit. The topics cohere. Probability, models, memes, institutions, and AI now appear inside one explanatory structure.

That may indicate reach. It may indicate internal consilience. It may indicate that the abstractions have been successfully reconstructed. It may also produce the very feeling of confirmation the chapter has been explaining.

The distinction cannot be resolved by adding more fitting examples.

The live question remains:

> Where does it break?

Anti-justificationism can sound like subtraction: no certainty, no induction, no probabilistic foundation, no confirmation. But the resulting epistemology is not empty. Far from it.

Problems still constrain inquiry because some attempted solutions fail to solve them. Explanations still assert dependencies that can contradict observations, other explana-

tions, or themselves. Models still have boundaries. Experiments still create situations in which failures can become visible. Measurements still reveal discrepancies. Statistical methods still characterize represented structures. Corroboration still records survival of severe criticism. Consilience still exposes conflicts between explanations that claim to describe one reality.

The rigor remains because error remains possible and because explanatory structure can be exposed to it. What disappears is the claim that rigor comes from accumulating support until belief becomes entitled.

AI changes the practical balance around that rigor. For most of history, large explanatory structures were expensive for an individual to carry. Authorities, institutions, texts, doctrines, and compressed cultural forms preserved more than one person could keep available at once.

Now a singular person, *you*, can increasingly interrogate explanatory surfaces that would previously have remained inaccessible. You can ask for the structure behind a slogan or the rival account behind a disagreement. You can reopen the model boundary behind an observation, the assumption behind a probability, or the unresolved criticism behind a confident summary.

That does not guarantee openness. The same system can produce a denser confirmation loop. But the available operation has changed.

The epistemic promise of AI may therefore be less about giving you better answers than about making it cheaper to keep more epistemic structure open to criticism at once.

This essay itself emerged through that kind of movement. It did not begin with a finished architecture.

It began with a problem:

> anti-justificationism conflicts with the felt naturalness of confirmation.

The first attempt could have stopped with a familiar answer: people seek certainty. That did not explain enough.

This essay was not justified into existence. It was an attempt to solve a problem, and the attempt remains provisional. Its reach may grow. Its consilience may grow. Neither protects it from your criticism. The next criticism is what could create the next problem.

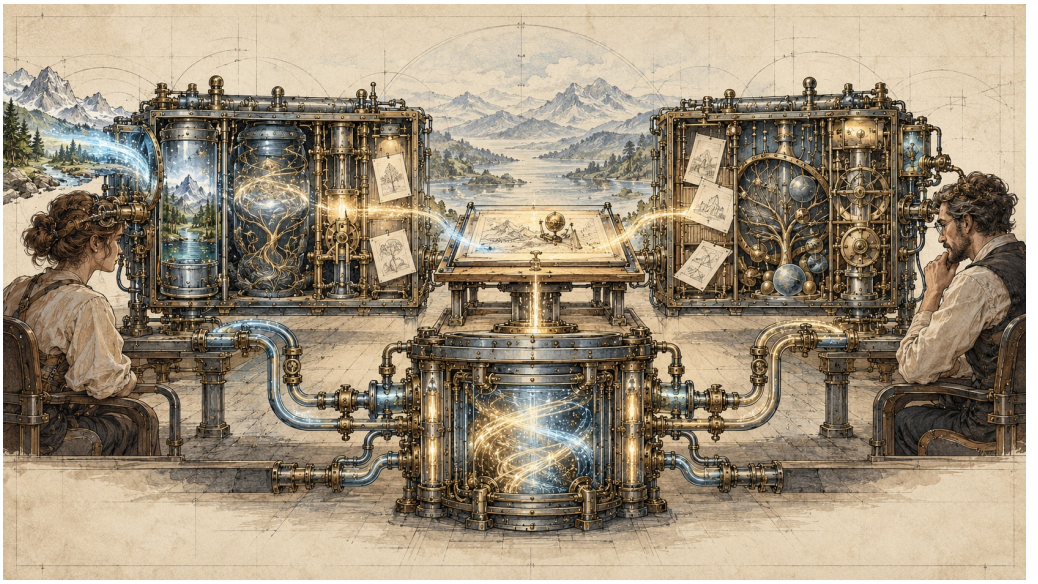
What is the next problem?

If criticism is a learned orientation, and if AI can make both justification and criticism dramatically cheaper, then the successor problem is no longer only how knowledge grows.

It is:

> How can you engage with objective knowledge while remaining open to criticism that might change the very problem you're trying to solve?

XIII. Gut Your Trust



We treat explanations as structures that can be compressed, transmitted, defended, criticized, preserved, and reopened. We distinguish explanations from models, models from experiments, and observations from the theories in which they acquire meaning. We argue that artificial intelligence can sometimes help you preserve more of an explanatory surface than unaided biological memory can comfortably carry.

But all of that assumed that the relevant knowledge had become explicit enough to appear on a surface at all. Much of what guides you never does.

A scientist may know that something is wrong before they can say what is wrong. A musician may hear that a phrase needs to move before they can explain why. A child may stop touching a hot stove long before they possess a theory of heat transfer. You may dislike a room, trust a stranger, avoid a street, return to a melody, or feel that a life decision is wrong without possessing an articulate explanation.

Something changed. Something was learned. But the knowledge is not yet explicit.

That suggests the next problem:

> **How much of your epistemic life exists before you can say it?**

And behind it sits a stranger one:

> **What if the machinery carrying that inexplicit knowledge differs radically from one person to another?**

You are inhabiting a unique, path-dependent epistemic mind. The next step is not merely that people believe different things. The machinery through which knowledge becomes available may differ too. Memory may or may not arrive as imagery. Thought may or may not be accompanied by inner speech. Sensory experience may be recalled through different forms, leaving different internal resources available for manipulating what you know.

The visible person gives you almost no direct access to this machinery. Similar behavior invites an inference of similar inner architecture. That inference can be profoundly wrong.

1. Before Knowledge Becomes Explicit

Start with the physical world.

There is one physical reality, and you are part of it. Explanations are attempts to describe structures within that reality. Physics gives us unusually rigorous examples: relativity, quantum theory, mathematical models that make parts of explanations explicit, experiments designed to expose implications to criticism, and observations produced under carefully constructed conditions.

Abstract it.

A simplified picture looks like this:

$$R \to E \to M \to X \to O$$

where R is physical reality, E is explanation, M is model, X is experiment, and O is observation.

Something important is missing.

The next explanation does not follow mechanically from the observation. A scientist notices that an anomaly matters. They feel that one interpretation is strained. They wonder whether two successful theories conceal a deeper contradiction. They imagine a critical experiment that no textbook yet contains.

Much of this may begin before the problem is articulate. A scientist may spend months with little more than:

> something here does not fit.

That is not yet a theory. It may not even be a fully formed problem. But it is already affecting attention, salience, and what can no longer be ignored. The problem is epistemically alive before it can be stated.

Put the scientist back into the picture:

$$R \to E \to M \to X \to O \to K_{\text{inexplicit}} \to P \to E'$$

Between observation and the next explicit explanation lies a living epistemic person.

The phrase *inexplicit knowledge* can sound mysterious. It need not. An animal learns a route to food. A child learns the timing of conversation before they can explain conversational timing. A pianist learns how much pressure a key requires without calculating force. A driver learns how close another car feels without being able to specify the geometry.

In each case, behavior changes in a way that reflects accumulated encounter with reality. The knowledge improves without first becoming a proposition.

So the conjecture

> knowledge must become explicit in order to improve

fails.

The simpler structure is:

$$K_{\text{inexplicit},t} \rightarrow \text{experience / criticism} \rightarrow K_{\text{inexplicit},t+1}.$$

No explicit statement need intervene. Organisms were learning long before any organism could state an explanation.

Then why does explicit knowledge matter?

Because once something becomes explicit, you can inspect it, state it, criticize it, compare it, teach it, write it, model it, preserve it, recombine it, and externalize it. So can other people.

An animal may learn a route. You can say:

> the lower trail floods after rain, so take the ridge.

Now the knowledge can leave the original learner. Someone else can criticize it. A map can modify it. Weather data can refine it. A future traveler can benefit without repeating the original experience.

Explicit knowledge does not create epistemic life. It changes the architecture and rate of possible criticism.

That difference matters because inexplicit knowledge is not merely explicit knowledge with the words missing. What is inexplicit may be a tension, a skill, a discrimination, a learned expectation, a pattern of salience, or a felt incompatibility. When language appears, it may create a new representation rather than transcribe a proposition that was already waiting intact inside you.

Call the process of making previously inaccessible knowledge explicit, **explicitation**. It is a creative act. You conjecture an explicit representation of something that was only partly available. The representation can be wrong. Once it exists on a surface, it can also be criticized.

That gives us the first bridge into the hidden world: knowledge can change before you can say what changed, and saying it does not merely reveal the old knowledge. It creates a new object on which criticism can operate.

2. Different Machinery, Similar Competence

Now keep the competence in view and vary the machinery.

Most of us carry an assumption we rarely state:

> other minds probably work more or less like mine.

We know that people differ in memory, personality, interests, and skill. We may imagine some as more visual, some more verbal, some more emotional, some more analytical. Yet the underlying cognitive machinery still feels broadly shared.

A simple question destabilizes that assumption:

> What happens inside your mind when you remember your childhood bedroom?

Do you see it? Do you know it without seeing it? Do you hear anything? Do you feel its spatial layout? Do you narrate the memory to yourself? Does the scene arrive from your own point of view, from outside, as a fact, as a feeling, or as a sequence of words?

There may be no single answer that represents the human norm.

Consider someone who cannot voluntarily visualize. Not dim imagery. None.

Ask them to picture an apple. There is no apple. Ask them to picture their mother's face. Nothing appears. Ask them to imagine walking through their childhood home. There is no internally projected scene.

If you possess ordinary or vivid imagery, this can be difficult to conceive. How can someone remember a face without seeing it? How can someone navigate without a map in the mind? How can someone rotate an object mentally if there is no object to rotate?

Yet people with aphantasia can navigate, recognize, reason spatially, remember, create, program, design, and plan. The outward deficit you might expect is often absent.

The relation that survives the case is simple:

> **the knowledge was never identical to the internal image.**

Someone can know what something looks like without consciously seeing it. They can solve spatial problems without consciously manipulating perceptual objects. The internal format differs. The competence can remain.

The surprise runs both ways. Someone with lifelong aphantasia may assume that *picture this* or *see it in your mind's eye* is metaphorical. Then they discover that other

people really do generate visual imagery, sometimes voluntarily and sometimes involuntarily. Some can also re-create sounds, voices, music, textures, smells, bodily sensations, and spatial scenes.

The person without imagery discovers that others may literally be doing something they had assumed nobody did. The visualizer discovers that a person can think and remember without machinery they had assumed was necessary.

Private phenomenology had been mistaken for universal architecture.

The contrast creates a conjecture. If conscious imagery is unavailable, does repeated reliance on semantic, symbolic, or relational structures make some forms of abstraction especially natural? Computer science offers an interesting case because programming often concerns invisible structures: relations, transformations, nested abstractions, formal constraints, and symbolic representations.

Perhaps some aphants become unusually comfortable there because they have long practiced cognition without conscious visual objects. Perhaps not. Nothing here establishes a category such as *aphants are abstract thinkers*.

What survives criticism is narrower:

> **different internal machinery may create different cognitive affordances.**

Move to the opposite pole. Imagine someone who can revisit previous sensory experience with extraordinary vividness. A childhood kitchen returns with the angle of afternoon light, the pattern on the floor, a voice from years ago, rain against a window, perhaps even the smell of the room and the bodily feeling of being there.

The obvious description is an extraordinary virtual-reality generator. The deeper move is to turn that description around. You already inhabit one.

This is not an analogy. Your nervous system is physically embedded in reality and interacts causally with it, but conscious experience is an internally constructed representation generated through those interactions. There is no second channel through which reality arrives in consciousness unrepresented.

You do not experience photons. You experience color. You do not experience molecular vibration. You experience sound. You do not experience neural firing. You experience distance, movement, warmth, texture, and significance.

Abstract it:

$$R \to S \to W_i$$

where R is one physical reality, S is embodied sensory interaction, and W_i is the experienced world constructed in person i.

The word *virtual* does not mean unreal. The representation is physically real. The person is physically real. The experience is physically real. But the experienced world is not an unmediated copy of external reality. It is a constructed interface with it.

Hyperphantasia makes this interface harder to ignore. If someone can return to a remembered moment with unusual intensity, a new set of questions becomes available. Does vividness give previous problem situations a stronger pull? Does sensory-rich memory feel more trustworthy merely because it is vivid? Can emotional events remain phenomenologically present long after their practical relevance has changed? Can rich imagery make internally generated material unusually available for artistic work?

These remain conjectural affordances, not destinies. One distinction is already secure:

$$\text{vividness} \neq \text{truth}.$$

A rich internal representation can still be mistaken, reconstructive, selective, or emotionally weighted. The spectacular quality of recall does not make it reality.

Now change another variable. Imagine someone whose memory and sensory recall seem broadly ordinary. They can visualize, remember, hear, and reason, but there is no continuous inner narrator: no voice explaining the day, no verbal stream announcing thoughts as they occur.

If you possess persistent inner speech, this can be as difficult to imagine as aphantasia is to a vivid visualizer. What happens when they decide? How do they know what they think? How do they plan, argue with themselves, or reflect?

Abstract it.

The erroneous identification is:

$$\text{thought} = \text{inner speech}.$$

The case breaks it. Thinking can occur without continuous verbal narration. A linguistic report may come later. A conclusion may become explicit only when communication requires it.

That creates another conjecture. If cognition is not continuously serialized into words, might some relations be held more holistically? Might a conclusion emerge before the verbal chain that explains it? Might language function more as output than as continuous internal workspace?

Possibly. Nothing here establishes an advantage. The explanatory gain is different:

> **the form in which thinking appears to consciousness is not identical to thinking itself.**

These cases do not establish three cognitive types. They expose dimensions that can vary independently: sensory imagery, memory re-experiencing, and inner verbal narration. Many other dimensions may vary too.

Can you hear music internally? Imagine smell? Summon bodily sensation? Do you remember autobiographical events as scenes or facts? Do thoughts arrive spatially?

Do you rehearse conversations? Do you experience memory from the first-person perspective or as an observer? Can you manipulate shapes without imagery? Do you need time before a thought becomes verbal? How much of your own problem situation is available to introspection at any moment?

The list is not the map. It is enough to show that the space is much larger than the familiar categories suggest.

3. Turn the Cases Back on Yourself

The cases become epistemically important only when they alter what you can infer about yourself.

You already know that your beliefs and memories differ from those of other people. The stronger claim is that the machinery through which your knowledge becomes available may also differ. The forms in which you remember, imagine, rehearse, discriminate, and explicate may be partly hidden even from you, because you have no independent view from outside your own implementation.

At first, your own mind feels normal. Then some people appear unusual. Then several dimensions begin to vary independently. Then those dimensions combine. Eventually the more plausible picture is that each person may instantiate an effectively singular cognitive configuration.

The movement resembles the expansion of a cosmos. Earth once seemed central because it was the standpoint from which the sky was seen. Your own phenomenology can feel central for the same reason: it is the standpoint from which every other mind is interpreted.

Once that assumption breaks, familiar surface categories become less explanatory. Someone needs long periods of rumination, and we call them slow. Someone seems emotionally distant, and we call them detached. Someone repeatedly revisits the past, and we call them obsessive. Someone prefers abstraction, and we call them analytical. Someone rarely narrates their inner life, and we call them unreflective.

Those descriptions may capture behavior. They do not yet tell us what form knowledge takes inside the person.

The explanatory question is causal:

> **what hidden epistemic architecture could make this pattern of behavior intelligible?**

That question reaches your own life as directly as it reaches your interpretation of others. Different forms of inexplicit knowledge may alter which problems become salient to you and which representations feel natural. They may affect what is easy

to manipulate, what becomes exhausting, what attracts prolonged attention, which experiences remain accessible, and which past situations remain active.

Abstract it.

A possible structure is:

$$K_i^{\text{inexplicit}} \to P_i \to I_i \to A_i.$$

Your hidden epistemic machinery shapes the field of problems that becomes available to you. That problem field shapes interest. Interest shapes action. Action changes the knowledge.

This does not explain why one person cares intensely about abstraction while another requires concrete sensory contact. It does not explain why one needs long periods alone before deciding while another thinks most clearly in rapid conversation. Nor does it explain why one seems unable to release the past while another appears almost discontinuous with previous experience.

But the failure matters. *That is simply their personality* had looked like an answer. Once hidden epistemic machinery becomes visible, it becomes a placeholder for a better problem.

For another person, the question becomes:

> **what hidden epistemic architecture makes this problem alive for them?**

For you:

> **what hidden epistemic architecture makes this problem alive for me?**

The same change in resolution alters introspection. Instead of asking only what you believe, you can ask what happens when you remember, when you plan, when a thought becomes verbal, when a problem becomes tractable, or when an old situation becomes present again.

Do you see anything? Do words appear? Does the thought arrive whole? Do you need time before you can say what you know? Does the memory feel present? What makes one problem easy to inhabit and another exhausting?

These questions are not diagnostic. They expose a variable that surface categories conceal: the form in which knowledge is available to a particular person.

Keep that variable separate from competence. Aphantasia does not imply poor spatial reasoning. No inner monologue does not imply weak self-reflection. Vivid imagery does not imply superior memory. Different implementations can produce similar outward competence. Variation in mechanism does not establish variation in epistemic kind.

Now the claim that a universe exists inside every person becomes less metaphorical. You may carry mathematics, music, family history, political theory, private humor, moral conflict, professional knowledge, imagined futures, and unsolved problems. But those worlds may not even be instantiated through the same phenomenological machinery from one person to another. The differences are not only semantic. They may be experiential, representational, temporal, sensory, and linguistic.

The visible person is therefore a narrow interface. You see behavior. You hear words. You observe choices. From those outputs, you infer an epistemic mind vastly larger than what the surface contains. The category may describe the interface. It does not explain the universe behind it.

That creates a tension. If the internal machinery can differ this much, what remains common enough for people to understand one another at all?

4. Communication Without Sameness

The first observation is easy to miss: people still communicate.

A person with no visual imagery can collaborate with a vivid visualizer. A highly verbal thinker can work with someone with no continuous inner monologue. A person with weak autobiographical re-experiencing can love someone whose memories remain intensely sensory. Across these differences, people still teach, negotiate, design, build, make music, raise children, run institutions, write science, argue, cooperate, and coordinate across continents.

The diversity does not make this less remarkable. It makes the mechanism more interesting.

Earlier we treated communication as reconstruction rather than mental-state transfer. One mind does not copy another. Instead:

$$M_A \rightarrow O \rightarrow M'_B$$

where O is an objective expression and M'_B is the recipient's reconstruction.

Hidden cognitive diversity sharpens the claim. The reconstruction may differ not only in content but in representational basis. A visual thinker may express an idea in words; a non-visual thinker may reconstruct the same relation semantically. Someone with inner speech may receive a diagram; someone without inner speech may convert it into a nonverbal relational structure.

Communication succeeds without internal sameness. The requirement is not identical representation. It is enough shared explanatory structure for the problem at hand.

Now look at the medium. A sentence is thin. A diagram is thin. A musical score is thin. A photograph is thin. A mathematical expression is thin. None contains the full internal world of the person who produced it. Yet another mind can reconstruct extraordinary

richness from these narrow channels. The medium is impoverished relative to the mind. The recipient supplies the rest.

This reverses the usual complaint about the signal. Human communication works not because the signal contains everything, but because the receiver can create explanatory structure from what the signal makes available. The receiver is a **universal explainer**.

Objective knowledge now returns with sharper boundaries. Speech, writing, drawing, notation, maps, photography, recording, film, and software are translation interfaces between private epistemic worlds. Each creates an objective structure that carries some portion of one person's internal world into a form another person can reconstruct.

The interface is always incomplete. That incompleteness is not simply an engineering defect, because the objective object was never identical to the private world that produced it. Its distinctive power is different: once knowledge is externalized, people who do not share the original person's internal implementation can criticize it.

Separate two operations that are easy to collapse: making knowledge explicit and making it objective.

Inexplicit knowledge can learn, adapt, guide, discriminate, orient, and evolve. It is the older layer. Animals were learning long before language, writing, mathematics, or formal science existed. Inexplicit learning is therefore not a secondary residue beneath explicit thought. Explicit explanation emerged later from this older, living epistemic machinery.

The scientist who cannot yet explain why an anomaly matters is not an exception. The same architecture appears when a relationship becomes troubling before you can say why. It appears when an artistic problem becomes active before its solution can be stated, or when an interest reorganizes your attention before you can explain its importance. The content changes. The epistemic structure may be continuous.

Abstract it.

When explicitation succeeds, a new loop becomes possible:

$$K_{\text{inexplicit}} \to O \to \text{criticism} \to K'_{\text{inexplicit}}.$$

The explicit object becomes a surface on which error can operate rapidly. Instead of waiting for repeated lived experience, you can imagine counterexamples, compare alternative explanations, seek criticism, manipulate models, and revise before acting.

Call that change in the rate and reach of criticism the **first great acceleration**.

A **second acceleration** occurs when explicit knowledge becomes objective. Once the objective knowledge leaves you, it can be preserved, copied, distributed, recombined, criticized by strangers, and recovered by future generations. Your solitary mind remains the place where knowledge is apprehended and where criticism matters, but the knowledge environment becomes cumulative.

One person's explicit surface becomes another person's starting point. That cumulative crossing of persons through objective knowledge is *civilization as epistemic recursion.*

Explicit knowledge is therefore not a superior substance placed above inexplicit knowledge. It adds inspectability, transmissibility, deliberate criticism, objective preservation, recombination, and cross-person comparison. It changes the architecture of possible criticism.

And once communication is understood as reconstruction across different implementations, a further possibility appears. What if the interface itself could adapt to the person rather than forcing the person to adapt to the interface?

5. When the Interface Adapts to You

Historically, communication technologies imposed strong representational constraints. To write, you had to translate yourself into language. To compose, into notation. To program, into code. To communicate visually, into drawing or photography. The person adapted to the medium.

That is enormously powerful. It also means that some internal forms map more easily onto available media than others. If your thinking did not arrive in the form the medium required, you had to perform more translation before anyone else could encounter it.

Multimodal AI creates a different possibility: the medium can increasingly adapt to you.

You can begin with speech, writing, images, sketches, sound, video, spatial description, or gesture. A system can help integrate those forms into a richer external representation. Another person can then receive the resulting structure through a different representational basis.

Abstract it.

For example:

$$\text{visual}_A \rightarrow \text{AI-mediated structure} \rightarrow \text{verbal}_B.$$

Or:

$$\text{spoken intuition}_A \rightarrow \text{diagram}_B.$$

Or:

$$\text{musical pattern}_A \rightarrow \text{formal representation}_B.$$

This is not direct transfer between minds. That remains impossible. It is still objective mediation. What changes is that the translation surface can adapt to the forms available at either side.

The same capability reaches backward into explicitation. You may know only:

> something about this is wrong.

You cannot yet say why. A system offers candidate distinctions. You reject some, recognize others, refine them, and eventually a previously inexplicit problem becomes explicit enough to criticize.

Do not make the machine's contribution mysterious. It has not discovered the true hidden sentence, because there may have been no hidden sentence to discover. It has helped construct a candidate surface.

That makes AI an *explicitation surface*: an objective space in which candidate representations can be generated, rejected, refined, and made criticizable.

You remain the problem-holder. The system can propose:

> yes, perhaps it is this distinction.

But you can still answer:

> no, that misses it completely.

The rejection changes the inquiry. Without it, adaptive translation becomes substitution.

The larger opportunity is therefore not to make unlike minds cognitively similar. It is to let unlike minds meet without requiring representational sameness.

Three earlier abstractions now converge. You inhabit a constructed experiential world. Communication requires reconstruction across private epistemic worlds. Objective knowledge supplies interfaces through which those reconstructions can become public and criticizable.

Multimodal AI can mediate all three. It can receive externalizations from one private world, transform them, and return them in forms more congenial to another. The machine does not eliminate solitude. It can make solitude more permeable.

Primitive versions of such adaptive multimodal interfaces already exist. The question therefore changes. It is no longer merely whether an adaptive interface can be built. It is what happens when such an interface becomes part of the ordinary route between private and objective knowledge.

That is where the same mechanism turns dangerous.

6. When Translation Becomes Control

Begin with an inchoate tension. The system supplies a familiar label:

trauma.

burnout.

narcissism.

avoidance.

anxiety.

The label may feel clarifying. It may also terminate inquiry too early.
Abstract it:

inexplicit tension → AI-supplied category → premature closure.

This is the old danger of *memetic compression* in a new form. A system can help make your hidden world easier to articulate. It can also overwrite that world with whatever explanation is culturally easiest to retrieve.

The difference is not whether the system produced a fluent representation. The difference is whether the new surface increases criticism or replaces it.

That distinction makes access to adaptive translation morally significant for a reason deeper than productivity. Better interfaces may change who can make themselves intelligible, who can receive knowledge in a form that fits their cognitive machinery, and who can expose their own representations to criticism.

Someone poorly served by conventional media may gain disproportionately. The technology can enlarge participation in shared knowledge. But that same fact creates the next set of problems: who has access, whose internal forms are treated as legitimate, which representations are normalized or flattened, and who is still required to conform to the interface.

An adaptive interface can normalize, classify, surveil, distort, compress, or manipulate. As it mediates more deeply between private worlds, its epistemic discipline becomes more consequential.

The relevant criterion follows from the distinction we already have. An interface enlarges epistemic access when it increases your capacity to express, criticize, and reconstruct knowledge without unnecessarily replacing your problem situation with the system's categories.

If the person loses the power to reject the representation, translation has already begun to become control. That power is not an ornamental safeguard added after translation; it is part of the epistemic mechanism that keeps substitution criticizable.

7. Common Humanity After Diversity

The diversity uncovered by these cases might initially seem to dissolve common humanity. Instead, it relocates it. The common feature is not one standard cognitive machine, one phenomenology, one memory system, one sensory format, or one inner narrator.

The commonality lies deeper. People can create problems, conjecture, reconstruct, compress, translate, criticize, coordinate, and build objective knowledge. The universality is epistemic. The implementation can vary.

This makes the familiar phrase *humans are smart apes with tools* explanatorily thin. The biological continuity is real. But the phrase compresses away the phenomenon that requires explanation.

You construct an immense internal world. That world can contain abstract structures with no direct sensory counterpart. Its representational machinery may differ radically from another person's. Yet you can still translate enough of it through objective knowledge to collaborate across those differences. The unexplained phenomenon is much larger than *smarter animal, better tools.*

Abstract it.

The full epistemic cycle now looks something like:

$$R \to W_i \to K_i^{\text{inexplicit}} \to P_i \to E_i \to O_K \to M_j \to K_j^{\text{inexplicit}} \to P_j.$$

Reality constrains experience. Experience changes inexplicit knowledge. Inexplicit knowledge shapes problems. Problems generate explicit explanations. Explicit explanations become objective knowledge. Other minds reconstruct them. The reconstructed knowledge enters different inexplicit machinery. New problems appear. No collective mind is required. The cycle crosses persons through objective knowledge.

AI enters this cycle as objective knowledge. It can help externalize, translate, preserve, compare, reformat, decompress, and explicate. You supply the problem situation. The machine supplies access to large-scale objective structure. Together, you can construct interfaces that neither biological memory nor static media could easily sustain. That is a genuine expansion of epistemic reach.

You are inside the experiment while reading this essay. The objective text is shared. The internal reconstruction is not. If you visualize strongly, you may have built scenes. If you do not, perhaps none appeared. If you possess inner speech, you may have heard parts of the argument. If you do not, the structure may have arrived some other way.

Yet communication still occurred. That is the phenomenon the earlier descriptions were trying to reach.

The uncertainty is not a reason to romanticize people. It is a reason to remain curious where the surface evidence does not close the explanation. The hidden world within is not an inaccessible mystical realm. It is the vast portion of ordinary epistemic life that

does not appear explicitly: skills, salience, expectations, memory structures, sensory forms, problem tensions, values, interests, and motivations.

Knowledge can move without minds becoming the same. You remain solitary. Reality remains one. The interfaces can improve.

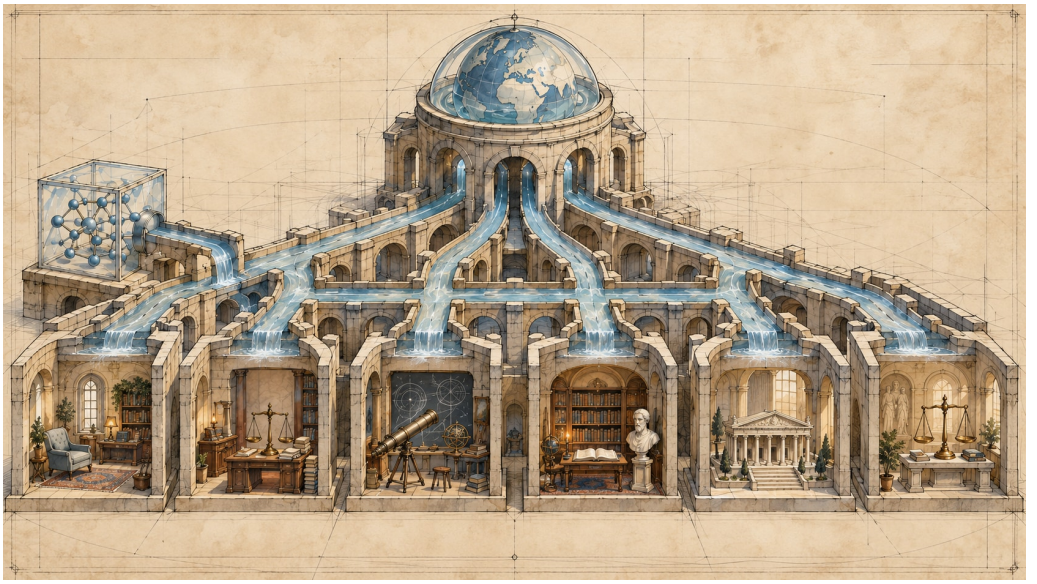
Once interfaces can adapt to the diversity of people rather than forcing them to adapt to the interface, the problem changes again. The question is no longer only what the machine can do. The new problem concerns what kinds of access between people can be built without collapsing difference into normalization. It asks which barriers can be reduced without deciding in advance which internal forms are legitimate. It also asks how your power to reject a representation can remain intact.

The hidden world within does not need to disappear for us to understand one another better. It needs better ways to become criticizable without being prematurely replaced.

That suggests a successor problem:

> **How can an adaptive epistemic interface help you make hidden knowledge criticizable without deciding in advance what that knowledge must become?**

XIV. Taking Explanations Personally



Inside your epistemic mind, we locate imagery, memory, inner speech, salience, skill, intuition, and other forms of inexplicit knowledge. What becomes explicit is only a narrow surface over an immense and largely inaccessible universe.

Making knowledge explicit gives your creative criticism something to chew on. But the explanation may still remain isolated from everything else you know. What happens when you allow its consequences to travel? What does it mean for you to possess an explanation, to really take an explanation personally?

Possession cannot mean merely repeating an explanation, assenting to it, defending it, or placing it beside other beliefs you also happen to hold. An explanation can have consequences outside the problem that first made it useful. It can connect to other explanations, rule out possibilities, expose contradictions, and create new problems. Call that capacity **reach**.

The difficulty is that almost everything in ordinary intellectual life can interrupt it. Knowledge is divided into fields. Professional jurisdictions are constructed. One question belongs to psychology, another to economics, another to physics, another to philosophy. Expertise can then acquire a second role beyond supplying specialized

knowledge: it can become a way of deciding where criticism is allowed to cross a boundary.

You can reproduce the same problematic structure inside yourself. An explanation can work in one part of your life while its implications are held away from another. Contradictions can remain dormant for years. An identity can remain protected from an idea you would readily apply elsewhere.

Until something stops fitting. A new fact makes an old belief uncomfortable. A moral judgment collides with something you know about another person. A theory you admire produces a consequence you do not want. The contradiction may still be inexplicit, but the world has become harder to make sense of. That experience may be one of the places where epistemology becomes embodied in the epistemic mind.

> **To take an explanation personally is to allow it to create problems for everything else you think you know.**

The claim is first epistemic. Its moral consequences arrive later, if you choose them.

1. One Reality Gives Explanations Reach

Under one reality, explanations that describe it may begin at different levels and solve different problems, but they cannot finally require incompatible reality. A theory of mind cannot require something physics excludes. A theory of knowledge cannot require a transformation that computation makes impossible. A psychological explanation cannot define people in a way that makes their known capacity for explanatory creativity impossible.

Explanations do not have to reduce to one another. They do not need the same vocabulary or the same object. They only have to remain mutually possible. That mutual possibility may be called **consilience**.

Now another consequence becomes visible. If an explanation describes structure in the one world, its consequences do not stop at the disciplinary border around the problem that produced it. They can travel. Evolution reaches beyond one species. Computation reaches beyond one computer. A theory of knowledge that claims universality reaches beyond philosophy and science into childhood, invention, institutions, morality, engineering, and ordinary life.

A problem in personality genetics generated a criticism of the transport from populations to persons. That exposed a social problem, then a problem about collective minds, then problems about freedom, abstraction, criticism, coordination, morality, justification, and inexplicit knowledge. The subject changed because the explanation kept producing consequences, and those consequences kept producing new problems. That is reach in practice; explanations traveling.

It also sharpens what it means to possess an explanation. You can say that people are fallible while treating one authority as incapable of serious error. You can say that individuals are free while making categorical exceptions when freedom becomes inconvenient. You can say there is one reality while protecting incompatible explanations inside separate fields.

In each case the explanation is still professed. What has been blocked is its ability to alter the surrounding problem situation.

Compare it:

$$\text{professing an explanation} \neq \text{taking an explanation seriously.}$$

Taking an explanation seriously means allowing what it explains to change what can remain unproblematic elsewhere.

2. Domains Are Not Jurisdictions

Specialization creates real differences in knowledge. A physicist may know things a psychologist does not. A surgeon may know things a patient does not. A statistician may see a modeling error everyone else misses. No individual can reconstruct all objective knowledge, so fields, technical language, professional communities, and specialized methods solve genuine coordination problems.

The problem begins when a difference in knowledge is turned into a protected boundary around criticism:

$$\text{specialized knowledge} \rightarrow \text{exclusive epistemic authority.}$$

Then a physicist is told not to criticize a psychologist, an engineer is told that a philosophical assumption in the system lies outside engineering, or an outsider's contradiction is dismissed because the outsider lacks the recognized credential.

The local knowledge may still matter. Expertise in cardiology does not supply the knowledge needed for aircraft maintenance. That is a difference in knowledge, not a difference in the kind of person allowed to expose a contradiction. Contradictions do not acquire professional memberships.

An explanation can therefore have a bounded domain of application without having a protected jurisdiction against criticism. The phrase "stay in your lane" can conceal the opposite idea: knowledge belongs to the people authorized to possess it. Under that picture, the expert possesses technical truth, the discipline possesses its object, and everyone else receives conclusions.

An inductive account of knowledge makes the move especially tempting. If knowledge is accumulated from observations, then the person with more observations appears to

possess more of the knowledge itself. The outsider has seen less, so the outsider can be classified as epistemically weaker before the criticism has even been reconstructed.

But once an explanation exists in a form that can be reconstructed independently of its original knower, biography no longer fixes its reach. A mathematical proof can outlive the mathematician who first understood it. A scientific theory can be reconstructed centuries later. A line of code can embody an explanation its current maintainer understands only partly.

This is one of the consequences of **objective knowledge**: knowledge can become available for reconstruction without remaining inside the person who created it. That does not make reconstruction automatic. It does change the question. The social identity of the critic cannot settle whether the criticism succeeds.

The relevant question is:

> What problem does the criticism expose?

3. When Reach Becomes Personal

The external silo has an internal analogue.

Picture your mind as a collection of beliefs: scientific beliefs, moral beliefs, beliefs about yourself, beliefs about other people, professional knowledge, family knowledge, private intuitions. That picture is much too simple. Much of your epistemic mind is inexplicit. It appears as expectation, salience, memory, emotional orientation, embodied skill, perceptual organization, habitual distinction, and tacit causal structure.

You do not merely hold propositions. You inhabit an epistemic mind. Your mind contains contradictions, but it cannot literally become several realities. It manifests your reality.

Compare it.

Let R represent reality and M_i the epistemic mind of person i. Then:

$$M_i \neq R.$$

Your epistemic mind is not reality itself. You do not possess complete, unmediated access to what exists. But M_i is not arbitrary either. You act in reality. Expectations succeed or fail. Objects resist. Other people surprise you. Predictions break. Tools work or do not.

There is therefore some continuing, fallible structural correspondence:

$$M_i \sim R.$$

The relation is not identity and not perfect representation. It is correspondence good enough, in some respects, for you to navigate the world and revise your understanding of it. This is the conjectured isomorphism.

Now the internal problem becomes clearer. You do not ordinarily experience M_i as a model displayed beside a separate reality. You walk into a room and experience the room. You see a face rather than consciously receiving uninterpreted wavelengths and then constructing one. When inexplicit knowledge about another person predicts danger, what appears in experience may simply be unease. Your epistemic mind is the medium through which reality becomes intelligible to you.

That is why a contradiction can first appear as a disturbance in intelligibility. A trusted explanation produces a consequence you cannot absorb. Someone you classified one way acts in a way the classification cannot explain. A principle you apply confidently in one setting becomes painful when it reaches your own child. A theory of yourself can no longer account for what you have done. The contradiction may not yet be explicit. Something simply feels wrong.

Criticism sometimes begins in felt disequilibrium. Inexplicit knowledge becoming available for explicit thought when a problem makes it relevant. Put those together and a further conjecture appears:

> **the pressure toward explanatory integration may be one way epistemology is instantiated in you.**

You may take knowledge seriously because, in some sense, you already live inside it. This is not a call for reductionism. You can love mathematics and music, be playful at home and procedural at work, and possess different forms of knowledge for different problems. Different identities, roles, habits, and forms of attention can coexist without difficulty. Plurality is not contradiction.

The problem appears only when two explanations make incompatible claims about the same reality. Then the incompatibility remains a problem even if it stays dormant or you would prefer not to integrate it. Multiplicity does not create multiple realities.

A stronger escape would be to declare some parts of personhood permanently outside explanation: the soul, love, art, intuition, intelligence, identity, mystery. The distinction here is not between explaining everything now and explaining nothing. It is between saying that no explanation is yet available and saying that no explanation is possible in principle. The first statement leaves a problem open. The second makes a claim about the permanent limits of knowledge.

That claim is itself explanatory enough to be criticized.

4. Expertise Supplies Criticism, Not Sovereignty

The Dunning–Kruger effect provides a small case in which these ideas converge. Its popular compression is that incompetent people do not know enough to recognize their own incompetence. There is a plausible mechanism inside that slogan: if you lack

knowledge, you may also lack the distinctions needed to notice some of your errors. Learning can reveal errors that were previously unavailable to you as problems.

The trouble begins when that mechanism is compressed again.

Abstract it:

$$\text{knows less} \rightarrow$$

$$\text{cannot evaluate} \rightarrow$$

$$\text{does not know that they cannot evaluate} \rightarrow$$

$$\text{their judgment has little epistemic standing.}$$

The framework here offers a different description. Suppose that at time t you possess knowledge K_t. That knowledge makes some errors visible and others invisible. It also supplies criteria C_t for evaluating performance:

$$K_t \rightarrow C_t \rightarrow \text{self-model}_t.$$

Then you learn. New distinctions become available. The criteria change. Errors that previously did not exist for you as explicit problems become visible:

$$K_{t+1} \rightarrow C_{t+1} \rightarrow \text{new self-model.}$$

No defective class of knower is required. Your current knowledge limits the criticisms currently available to you. Expertise matters because an expert may possess criticisms you cannot yet formulate. If an engineer identifies a structural dependency in a bridge design that you do not understand, your ignorance is real.

But the difference is:

$$\text{knowledge possessed,}$$

not:

$$\text{kind of knower.}$$

You can learn. The expert can be wrong. An outsider can expose a contradiction. Unequal knowledge can explain unequal access to criticism without creating an exclusive jurisdiction over truth.

That makes other people epistemically important for a different reason. If your epistemic mind is partial with respect to a subject (the universal state of ignorance), another person may carry an observation you lack, an explanation you have never encountered, or a problem your abstractions do not make visible. They may carry precisely the perturbation your epistemic mind cannot generate from inside itself. But the perturbation is useful only if you recover what it means.

Suppose another person says "AI is unsafe," "experts know better," or "children need discipline." Each sentence is compressed. It may carry observations, causal claims, moral judgments, inherited ideas, personal history, vague unease, or an explanatory criticism. The surface sentence does not tell you which.

The first operation is therefore to open the compression and conjecture the structure that produced it. The other person can then correct that reconstruction before criticism begins.

Call this operation **decompression**:

expression → decompression → conjectured meaning → correction → criticism.

This differs from steel-manning. If the other person produces explanation E_A and you replace it with a stronger explanation E_B, you may have improved the debate while losing the thing they actually meant. Decompression does not improve the other person's explanation. It attempts to recover it. Steel-manning often does the opposite: it strengthens the compression, making it more resilient to criticism.

Communication can't give you direct access to another mind. Recipients reconstruct meaning; two reconstructions can be produced through radically different internal machinery. Taking another person's explanation personally therefore cannot mean accepting it. It means reconstructing enough of its content that the explanation can actually be criticized.

Once that happens, an objection changes status. It is no longer merely an obstacle to your explanation. It may contain the missing problem. The other person may still be confused or wrong. The point is that social classification no longer decides the matter in advance. That is the form of criticism needed here: criticism creative enough to let another epistemic mind change your problem.

5. Many Tools, One Epistemological Problem

A more serious objection now appears. Perhaps there is no single epistemology at all. Perhaps there are only different ways of knowing: induction, Bayesian inference, falsification, deduction, abduction, intuition, testimony, decision theory, statistics, common sense. Different problems call for different tools, and rationality consists in choosing the right one.

That picture is attractive because it looks tolerant and practical. It also resembles ordinary scientific and technical work. Researchers use statistics. Engineers use models. Doctors reason from evidence. People update beliefs. And of course, a Popperian conjecture is guesswork.

But the existence of many epistemic operations does not yet establish many epistemologies. Bayesian inference has broad applications. Statistical estimation is com-

monplace. Deduction is indispensable. Models and observations matter. Heuristics can generate candidate solutions. Intuition can be the first explicit surface of an inexplicit disturbance.

The objects are different. An explanation asserts structure. A model represents some of that structure for a problem. An experiment creates a critical situation. An observation occurs within that situation. Probability distributes weight across a represented possibility space. A heuristic proposes a move. None of those descriptions makes the operations unreal or unimportant.

They can still participate in one broader process:

$$P \rightarrow TS \rightarrow EE \rightarrow P_2.$$

The pluralism may therefore be functional rather than foundational.

Bayesian reasoning makes the distinction concrete. Suppose a gambler knows the rules of a game, the deck is specified, the outcomes are represented, and the relevant probabilities can be calculated. Probabilistic reasoning may be exactly the right tool. The present epistemology has no reason to deny that.

The problem begins only when the tool is asked to explain the creation of the represented space in which it operates. Probability can redistribute weight among available possibilities. But a creative problem may consist in discovering that the variables are wrong, that the relevant hypothesis does not yet exist, that the evaluative criterion needs to change, or that the problem itself has been mis-stated. Probability can operate brilliantly once those structures exist. That does not yet explain where they came from.

The same difficulty appears if rationality is defined as choosing the right tool. Use Bayes when Bayes applies. Use deduction when deduction applies. Use intuition when intuition is useful. The instruction merely moves the explanatory burden into the word *right*. What makes one operation appropriate when the problem itself can change?

What unifies the cases here is not a common surface procedure. It is a direction. A musician can notice an error inexplicitly. A physicist can derive a contradiction. A statistician can expose a model failure. Different operations can all alter the problem by exposing error.

That direction is what this epistemology calls **rationality**:

Rationality is an orientation toward exposing and eliminating error.

The visible tools can differ while the epistemological problem remains the same.

6. Universality Means Vulnerability

The epistemology explored here does not present as one useful perspective among others. It claims that knowledge grows evolutionarily, through problems, conjecture,

criticism, error elimination, and successor problems. That claim is universal. So it is vulnerable wherever knowledge grows by an incompatible mechanism.

Epistemic pluralism therefore cannot be dismissed as a matter of taste. It is a direct criticism of the universality claim. The anti-inductivist part is especially exposed. If explanatory knowledge can actually be logically derived from repeated observations, or more generally, if knowledge may be derived from data, then a central part of this epistemology is not merely incomplete. It is false.

Abstract it.

The break condition can be stated directly. If there exists a valid operation:

$$O_1, O_2, \ldots, O_n \rightarrow E$$

in which observations logically generate the explanatory content without a creative conjecture, then the account of knowledge creation offered here has failed. Redefining that operation after the fact so that it fits the theory would only protect the theory from the consequence of its own universality.

Metaphysics does not remove this exposure. Epistemology concerns the structure of knowledge and its relation to reality, so an isolated observation cannot stand outside all explanation and interpret itself. A measurement already sits inside a model, an experiment, a theory, and a problem situation. But that does not make epistemology immune from criticism. A rival explanation with greater reach can expose a contradiction here. Engineering can reveal a knowledge-creating mechanism this account excludes. A supposedly impossible operation can become a new problem for the theory.

The relevant standard is not whether one observation can mechanically refute metaphysics. It is whether the explanation survives contact with everything else we know.

Beliefs and credences occupy the same position. A model can begin with beliefs, update credences when evidence arrives, and derive actions. That may solve real problems. But the present explanation of knowledge growth does not require every epistemic change to pass through an explicitly represented degree of belief. You can act on a tentative solution without calculating a credence. You can discover a problem that was not represented, change an evaluative criterion, or construct an explanation that did not exist among the prior options.

Conventional concepts of belief or decision-making do not acquire foundational status merely because they are conventional.

The openness runs both ways. If a future theory of belief or decision acquires universal explanatory reach and collides with this account, the collision becomes part of the problem situation. No explanation is immune. Including this one.

7. Engineering Must Reach Both Ways

Engineering now becomes the surface for a severe test because epistemology has crossed into a domain where explanations can be embodied in artifacts.

If knowledge creation requires open problems, conjectural variation, criticism, revision, and the possibility of changing evaluative criteria, then an architecture intended to create knowledge provides a place where those claims can acquire technical consequences. If the explanation has no consequences at all for such an architecture, its supposed reach has stopped exactly where the test becomes concrete.

But the direction cannot run only from philosophy into engineering. Engineering creates new objective knowledge. Systems fail in unanticipated ways. Architectures surprise their designers. Capabilities appear that were not predicted. Mechanisms that seemed necessary can turn out to be incidental. If an engineered system creates knowledge through a process incompatible with this epistemology, then the engineering has created a criticism of the epistemology.

Abstract it.

The relation is reciprocal:

$$\text{epistemology} \rightarrow \text{engineering constraints}$$

and:

$$\text{engineering} \rightarrow \text{criticism of epistemology.}$$

Neither side owns the other.

The conversation that produced this work adds a recursive case. The artificial system participating in the work has been deliberately configured, or harnessed, around the epistemology under discussion. It represents inquiry through problems, tentative solutions, error elimination, and successor problems. It is constrained against treating authority, confidence, consensus, or supporting evidence as epistemic foundations. It is asked to expose contradictions rather than merely defend outputs. It also distinguishes the human operator and originator of the problem situation from the machine's role in retrieval, recombination, formalization, and execution.

That arrangement proves nothing about the epistemology. A system can be built according to a false explanation. The significance is different. If an explanation claims consequences for invention, one way to expose those consequences to criticism is to instantiate them and see what breaks. This system is one such attempt.

Its characteristic failures would therefore matter. A formal cycle of problems and error elimination might suppress useful creativity. The system might fail to originate problems without human intervention. The distinction between conjecture and induction might become impossible to maintain in implementation. Enforced criticism might collapse

into ritualized self-critique. An architecture designed for explanatory openness might become procedurally rigid.

Those failures would be more than defects in a tool. They would be observations created by embodiment: criticisms that were unavailable while the epistemology remained purely philosophical. That is what reciprocal reach looks like. The explanation constrains the artifact; the artifact gets a chance to break the explanation. Reality kicks back.

8. Abduction as a Test of Creative Knowledge

One proposal at the frontier of artificial intelligence sharpens the engineering test. Current systems may be extraordinarily capable statistical and inductive machines, yet still lack abduction: inference to the best explanation. Perhaps that missing operation is the bridge to genuine explanatory creativity.

Abstract it.

The proposal has a familiar form:

$$\text{observations} \rightarrow \text{candidate explanations} \rightarrow \text{best explanation}.$$

If that process genuinely explains open-ended explanatory creativity, then the epistemology explored here has a serious problem.

The pressure appears inside the phrase *best explanation*. Best according to what? Reach, simplicity, likelihood, fit, prior probability, utility, or some learned objective? Whatever the answer, the architecture requires criteria by which alternatives can be compared.

But there is a much harder problem. A creative solution may not merely satisfy the criteria available before the solution. It may alter what the problem is and what counts as a good answer.

The structure can therefore change from one problem-and-criteria pair to another:

$$(P_1, C_1)_t \rightarrow TS \rightarrow (P_2, C_2)_{t+1},$$

where C_i represents evaluative criteria. The map changes with the destination.

That temporal asymmetry matters. After a creative leap, the successful explanation reorganizes the observations, exposes weaknesses in earlier theories, and may look like the obvious best explanation. Retrospectively, the history can be compressed into a clean sequence: the thinker observed the evidence and inferred the best explanation.

But the later description may erase the event that needs explaining. Before the explanation existed, the relevant hypothesis was not necessarily one member of an already represented set. Before the problem was reformulated, the criterion by which the later answer looks best may not have existed either.

Calling the process abduction can therefore restate the finished achievement without yet explaining the origin of the new hypothesis, problem, or evaluative criteria.

That criticism does not imply that abductive methods are a dead end. Quite the contrary. They may improve hypothesis ranking, organize observations into more useful candidate explanations, expose latent relationships, strengthen scientific tools, and generate outputs that increasingly resemble explanatory reasoning.

Current induction engines may become better induction engines.

The conjecture here is narrower:

> **abduction, if it remains inference within inherited or learned problem and evaluation structures, will not by itself produce the open-ended explanatory creativity attributed here to people.**

And further, where creative variation in this strong sense is introduced, we will trace the origins to people. That conjecture has to remain breakable. Otherwise the conjecture would define away every contrary result.

The framework attributes several capacities to people: originating new problems, generating explanations not logically derived from prior observations, creatively criticizing those explanations, transforming evaluative criteria, and creating successor problem situations as novel structures. If an abductive architecture produces those capacities, then the epistemology has a serious problem. It doesn't yet have a rival explanation, but a serious problem.

This posture sits comfortably, if not obviously, with the knowledge that inductive *inference* does not exist (in the logical sense), and the much older observation that people conjecture in the inductive *direction*, and the engineered reality of systems that process data inductively (in a mechanical sense as engineers describe it). This may be confusing, but it is not confused. We don't argue about words, we decompress them.

AI therefore occupies an unusual position. These systems can produce surprising outputs, recombine enormous bodies of objective knowledge, surface analogies, generate candidate explanations, help expose contradictions across fields, simulate criticism, and preserve explanatory surfaces larger than you can hold in working memory.

Those capacities still leave the personhood problem open. Remaining questions may concern the active problem for the machine and where that problem came from. They may also concern what makes one criticism matter rather than another, and whether the system can transform the criteria by which its own explanations are evaluated. And if so, how other universal explainers may recognize the latent evaluative criteria. The questions are open because the engineering has not yet dissolved these distinctions.

At this point the explanation turns on itself, and you. Taking explanations personally applies to psychology, physics, engineering, probability, AI, other people, your own

internal knowledge, and the universal epistemology that connected them. Universality does not create a privileged final layer. It increases the surface for further criticism.

9. Moral Knowledge Enters the Same World

The moral consequence arrives only after that epistemic structure is in place. Moral explanations concern problems involving persons, suffering, freedom, responsibility, coercion, flourishing, and action. If they are explanations, they inhabit the same reality as every other explanation. Their consequences can therefore collide with what we know elsewhere, and criticism can expose those collisions.

That is the sense in which moral seriousness is explanatory seriousness. The claim does not make an authority true because the subject is morally important, and it does not insulate a moral explanation from criticism. It places moral knowledge inside the same consilient reality.

The familiar *is–ought* problem immediately creates pressure. An explanation of reality does not entail a moral obligation. So what kind of action-consequence can an explanation generate without smuggling an authority into the premises?

Return to the gambler. Suppose the active problem is to maximize expected return under specified rules, and the probabilities are known. The calculation can identify a current solution to that stated problem. You remain free to choose otherwise.

A contradiction appears only if all of the explanatory structure is held fixed while the opposite action is also described as the better solution to the same problem. If the problem changes, the action can change. If the probabilities are rejected, the calculation changes. If another value enters, the problem was not the one first stated. If some hidden condition matters, the explanatory structure was incomplete.

Something has to move. That distinction gives a precise role to the phrase **explanatory ought**. A coercive ought derives force from an authority: *you must do this because the authority requires it.* An explanatory ought does something different: *given this problem and this explanation, this action currently functions as a tentative solution.*

The second statement does not entail obedience. It exposes a consequence inside a problem situation. And people may disobey epistemically by altering that problem situation. You can reject the problem, criticize the explanation, introduce another value, construct a different solution, act inconsistently, or accept the consequences. Freedom remains because explanatory structure does not remove your capacity to create a new knowledge.

Earlier claims became morally relevant without being converted into oughts. If people are universal explainers, institutions that classify some people as epistemically inferior assume an explanatory problem. If criticism is required for error correction, structures that prohibit criticism acquire another. If objective knowledge can coordinate

heterogeneous people without merging their minds, claims that deep conformity is necessary for social order become criticizable. If children possess less knowledge without possessing a lesser kind of mind, authority relations built on the opposite assumption collide with a theory of personhood. If coercion closes problem situations, new coordination technologies can alter the conditions under which older coercive arrangements once semmed necessary.

Epistemology does not deduce a moral code from facts. It causes some moral explanations to become inconsilient with existing explanations and authorities.

Moral progress can then be described as a change in explanatory integration. A practice may once look necessary or harmless. A better explanation of personhood, a newly visible form of suffering, or a technology that removes an old coordination constraint can make the old explanation harder to sustain. Criticism from people subjected to the practice can expose a problem the earlier description suppressed. The world of moral knowledge reorganizes because its internal explanatory structure changed.

That creates one more danger. "Take explanations seriously" can itself be turned into an authority: *the science is settled, therefore obey*, or *this theory has consequences, therefore compliance follows*. That reverses the argument. Explanatory seriousness increases exposure to criticism; it does not convert an explanation into a command.

The strongest consequence of taking an explanation seriously may still be the criticism that destroys it.

10. Consilience Becomes Personal

The same structure changes what it means to take another person seriously. Their visible exterior must be only a compression of a vast epistemic mind. Taking the person seriously therefore does not mean assuming they are right or suspending criticism. It means reconstructing the problem, observation, inherited idea, causal structure, or unease that the expression is carrying far enough for criticism to reach the thing they actually intended.

That is the interpersonal form of decompression developed earlier. The person can reject your reconstruction. Only then does the remaining disagreement become a useful object of criticism. The point matters because the epistemic mind inside you was never purely private in the first place.

Imagine living almost entirely alone while possessing a modern scientific worldview. Your internal knowledge would still contain language, mathematics, physical explanations, inherited moral concepts, technologies, and theories created by people you never met. Isolation would not create a different reality. It would leave you densely connected to objective knowledge produced by others.

Even solitary thought is socially and historically populated. Language carries distinctions inherited from other minds. A scientific worldview carries centuries of criticism. A programming language embodies explanations about computation and abstraction. You reconstruct these structures locally, but they remain part of your epistemic mind. The relation between M_i and R is therefore mediated not only by your direct encounters with physical reality but by enormous bodies of objective knowledge.

You carry fragments of civilization inside your epistemic mind. The external and internal arguments now meet. Explanatory enclosure is artificial not only because contradictions cross disciplinary boundaries, but because the knowledge inside you already crosses them. The lanes were crossed long before anyone told you to stay inside them.

This coming together of knowledge is called **consilience**. Here the same idea becomes personal. Consilience is not only the condition in which physics and biology remain mutually possible. It is also the provisional intelligibility of your own experienced world. Your scientific explanation does not destroy your theory of personhood. Your moral explanation can coexist with what you know happened. Your memory and your account of yourself can still occupy one reality.

For a time, the system is in explanatory equilibrium. Then criticism arrives. Something stops fitting. The equilibrium breaks. A problem becomes visible. Knowledge changes. Taking explanations personally is not an extra rule added beside consilience. It is the process by which consilience remains open rather than hardening into isolation.

11. Maximum Exposure

The growing coherence produces a danger of its own. Genetics, psychology, society, freedom, abstraction, criticism, coordination, morality, justification, inexplicit knowledge, and now explanatory consilience have been forced together. As the connections accumulate, the framework can begin to feel inevitable.

That feeling is not a warrant. Internal coherence is not proof. Reach is not justification. Consilience is provisional. A framework with greater reach also has more surfaces on which contradiction can appear. Its success therefore increases the number of places where it can fail.

Some of those break conditions are already visible. If induction can create explanatory knowledge from observations without a creative conjecture, the epistemology is in jeopardy. If probabilistic updating can explain the open-ended creation of new problem situations and evaluative criteria, the epistemology is problematic. If a theory of belief or decision acquires universal reach incompatible with this account, the collision has to be resolved. If engineering creates an artificial person through mechanisms the epistemology excludes, the epistemology is denied. If moral explanations require a

fundamentally different account of knowledge from every other domain, the universality claim is false, and so is the epistemology.

These are not concessions added to make epistemology sound cautious. They follow from the claim that explanations acquire consequences outside the place where they began.

Now apply that claim to itself.

> Take seriously that *you* should take explanations seriously.

The sentence becomes meaningful only if it is allowed to create problems for every other explanation you hold. Rival explanations, engineering results, moral consequences, internal contradictions, and the architecture of this project all become possible sources of criticism.

If the recursion exposes a contradiction, the next problem has appeared. That is the epistemic cycle operating, not an exception to it.

Nothing in this requires omniscience or infallibility. Contradictions can remain inexplicit for a long time. Expertise can still supply criticisms you lack. Statistical and probabilistic tools can still solve real problems. Two people can reconstruct each other accurately and still disagree. Explanations can expose consequences without commanding obedience.

The requirement is narrower:

> **do not protect a good explanation from its consequences.**

The movement is now visible, but the mechanism is not. Why does one contradiction become irresistible while another remains compartmentalized? Why does one criticism transform your evaluative criteria while another fails to enter the problem at all? How does an inexplicit mismatch become explicit? How can a physically instantiated epistemic mind reorganize while preserving the continuity of the person who experiences the change?

Those questions are now harder because the proposed answer cannot be merely psychological. The mechanism has to remain compatible with this evolutionary epistemology, with computation, with physical instantiation, and with the personal continuity of *you*.

If taking explanations personally is one way epistemology is instantiated in people, then we still lack an explanation of how you do it. That successor problem makes explicit the question that has been waiting underneath all along:

> **What is a person?**

XV. The Recursive Universal Explainer



Taking explanations personally means allowing what you think you know to disturb what else you think you know. When that disturbance creates a new problem, it reinstantiates what knowledge-creating people have been doing all along: recursively explaining.

One compression has done enormous work:

the universal explainer.

It has distinguished people from genes, traits, populations, institutions, technologies, and other animals. It has advanced the argument that you cannot be reduced to the classifications applied to you. It has helped explain why freedom matters, why criticism matters, why objective knowledge matters, and why your future cannot be inferred from your present state.

But the phrase has become a meme. That does not make it false. Some memes are merely compressed slogans. Others represent an explanation so effectively that the compression begins to look like the explanation itself.

Universal explainer may stand close to the truth, entailed by our best explanations. But if we take it personally, we now have to decompress it and let it run. Understanding

explanatory universality as a consequence of computation and physics is to take it seriously. Exploring it as instantiated in an epistemic mind is to take it personally.

What kind of thing is capable of universal explanation? What must be true of you for that capacity to exist? How can an evolved biological organism create explanations about domains that had nothing to do with the environment in which it evolved? How can a finite physical system possess no known domain-specific ceiling on what it can understand? What roles do computation, abstraction, objective knowledge, and tools play in making that universality practical today, and what might it suggest about the machines we could build tomorrow?

This is not an attempt to define personhood once and for all. It is an attempt to let those explanations reach the person doing the explaining.

1. The Universal Explainer

Humans are small animals on one planet in an enormous universe. We are fragile and frequently wrong. Yet the same organism can explain stellar evolution, the structure of DNA, and mathematical structures no one can perceive directly. It can build instruments that operate beyond unaided sensation and then criticize the epistemology by which those instruments are understood.

The unusual capacity is not merely difficult problem-solving, adaptation, or learning. It is the creation of explanations whose reach is not confined to the environment in which the organism evolved. That is the claim compressed by *universal explainer*.

Universality does not mean that you already know every domain, or that every problem is practically solvable. It means that no known domain-specific boundary follows merely from the fact that you are a person. Given sufficient knowledge, criticism, tools, time, and physical resources, your explanatory reach is not restricted to some preassigned domain of problems.

Among the animals we know, nothing else appears to possess this open-ended capacity. Other animals learn, solve problems, communicate, remember, navigate, use tools, and coordinate socially. But we do not presently know of another animal that creates open-ended explanatory knowledge. So *universal explainer* is not an empty honorific. It identifies something objectively exceptional.

The danger is to let that unusual property become the whole ontology of the person. That would repeat the categorical error: a property of you, however important, is not a complete explanation of you.

The phrase therefore suggests a problem. If universal explanation is a real capacity of people, then explanations that apply everywhere else, including epistemology, must also be allowed to reach the entity doing the explaining.

Evolution reaches you. Physics reaches you. Computation reaches you. History, embodiment, and error reach you. Being able to explain them does not place you outside them.

2. Evolved, and Still Evolving

Begin with the obvious consequence. Human beings evolved.

No explanation of personhood can require otherwise. Yet you can understand evolution perfectly well in the abstract while preserving an inexplicit picture of yourself as something fundamentally outside it. In that case the explanation has not yet travelled all the way to you personally.

You are an animal. Your body is constructed through biological processes. The brain in which explanatory knowledge becomes physically instantiated is an evolved organ, and the sensory systems through which you encounter reality are evolved systems. Memory, affect, attention, perception, pain, pleasure, attachment, motor control, sleep, and development all arise within that evolved architecture.

Nothing about explanatory creativity requires an exemption from this history. The striking fact is the opposite: evolution produced an organism capable of explaining evolution.

But *evolved* points backward. It names the history that produced the present organism. The epistemology explored here has never treated you as complete. Knowledge changes what becomes salient. A new explanation creates new problems. A criticism makes an old abstraction visible. A skill changes what you can notice. A relationship changes the problem space. A technology makes previously unavailable forms of action possible.

So you are not only evolved. You are evolving. Not only genetically, but also epistemically.

$$K_t \rightarrow P_t \rightarrow TS_t \rightarrow EE_t \rightarrow K_{t+1}.$$

Variation, criticism, retention, and path dependence remain active. Your existing knowledge becomes part of the environment in which your future knowledge develops. This matters because creativity can look like the opposite of continuity. You can suddenly see a problem differently. A scientific explanation can reorganize a field. A work of art can alter what later creators perceive as possible. A moral insight can make an inherited practice intolerable.

The explicit change may be discontinuous. The causal system in which it occurs is not without history. Your mind depends on biological architecture, memory, language, skill, social knowledge, tools, habits, institutions, and objective knowledge. A new explanation enters that already complex structure. Some existing knowledge survives. Some becomes contradictory. Some tools or institutions can absorb the change; others

have to reconstruct around it. Some habits outlive the explanation that once supported them.

Compare it.

The relevant distinction is not:

incrementalism versus creativity.

It is:

creative discontinuity in explicit explanation

within

continuous causal systems that must reconstruct around it.

This is why piecemeal improvement can be more corrigible than wholesale reconstruction. When many dependencies change at once, the causal source of failure becomes harder to identify. Radical change remains possible, but error is easier to localize when reconstruction proceeds through sequences whose dependencies can still be criticized.

The same applies to you. Open-ended evolution is not instantaneous. Bodies, skills, relationships, concepts, and inexplicit knowledge all have histories. You can become radically different without being able to discard every inherited structure by choice. Your future is open and causally contingent.

3. Computation, Abstraction, and Tools

The next consequence begins with physical instantiation. Whatever happens when you remember, imagine, explain, criticize, love, compose, suffer, calculate, recognize a problem, or change your mind is realized through physical processes. If your explanation of computation correctly constrains the transformations physical systems can perform, then personhood cannot depend on a transformation that simply escapes those constraints.

That is where computation reaches the person. Calling a person computational can sound reductive because the word evokes machinery, automation, determinism, and lifeless symbol manipulation. But the explanatory claim is narrower: If human explanatory creativity requires a non-computational physical process, then that process itself requires an explanation under the physics that permits the person to exist. The discomfort does not remove the problem.

Computational universality has a further consequence. If a physical process can in principle be simulated by a universal computer, then some program can perform that simulation. That program is objective knowledge. The program in question may

be enormous. It may require impractical hardware, more time than a human lifetime, or more energy than can realistically be supplied. Those limits matter, but they are engineering limits rather than principled boundaries on what can be represented or computed.

Compare it.

The distinction is:

possible in principle ≠ practical under present resources.

No finite person can perform every possible computation unaided. Neither can any particular computer. Universality means that a change of domain does not by itself require a fundamentally different kind of computing machine. Given the right program, time, and resources, the universal machine can perform the relevant computation.

That matters because explanatory universality has never been realized through biological memory alone. People create objective knowledge and embody it in symbol systems, instruments, and technologies. Writing and mathematical notation preserve structure outside biological memory. Instruments extend perception. Computers, networks, and artificial intelligence make new transformations of objective knowledge practically available.

The finite biological person is therefore not the same explanatory unit as the unaided biological processor. Abstraction is part of what makes that extension possible.

A finite system rarely needs a complete microscopic simulation in order to explain or manipulate a higher-level structure. Software engineering makes the relation ordinary. A programmer can work with a string without continuously reasoning about transistor physics. An operating system hides much of the hardware. A runtime or language hides lower-level operations, while libraries and applications hide repeated operations again. The higher level is not independent of the lower one. It depends on it while suppressing detail that does not matter to the current problem.

A useful abstraction can therefore be represented roughly as:

lower-level causal structure → problem-relevant compression → higher-level manipulation.

The compression can fail. A timing bug can make concurrency relevant. A memory problem can expose an assumption hidden by a runtime. A hardware fault can propagate upward through every software layer. The abstraction was not unreal. It was locally adequate until an ignored detail became causally relevant.

Human explanation may work the same way. We do not calculate quantum states to explain why a bridge carries a truck. We do not simulate every neuron to understand why a criticism changed your mind. We do not reconstruct molecular dynamics to reason about a contract. Different explanatory levels preserve different causal dependencies.

That is one reason you can reach so far: explanation need not contain every lower-level detail in order to preserve the structure relevant to the problem at hand.

Now the problem sharpens. Physical processes may be computationally simulated in principle, while abstractions let finite systems manipulate progressively deeper structures without reproducing every detail.

Given both claims, what establishes a principled upper boundary on explanatory reach? Resource limits establish practical boundaries. A person has finite time. Civilizations have finite energy. Computers have finite memory. Some computations may remain impossibly expensive for any realistic purpose. But those constraints do not by themselves explain why some domain of reality would be forever beyond explanation in kind.

The stronger boundary claim needs its own explanation. Saying that people did not evolve for some domain does not establish that boundary; humans did not evolve specifically for general relativity either. Saying that the unaided brain cannot directly perform a computation does not establish it; much modern calculation already occurs in machines whose operations exceed what biological cognition can perform directly.

You create tools. The tools carry objective knowledge. That objective knowledge makes new transformations practically available.

Abstract it.

A more realistic unit is therefore:

$$\boxed{\text{You}\Big(\text{Tools}\big(\text{Objective Knowledge}\big)\Big)}$$

The nesting matters. You create objective knowledge. Some of that knowledge becomes technology. The technology enlarges your problem situation. The enlarged problem situation permits new explanations, which become new objective knowledge and can be embodied in further tools.

The process loops. That lets the existing phrase travel. You are a **universal explainer**. The noun names a capacity made possible by a history of incremental changes. You do not possess universal knowledge. No person has completed explanation or contains the content of every science, art, language, history, or possible future theory. So the verb matters too. You are **universally explaining**.

The parallel with evolution is useful:

$$\text{evolved} : \text{evolving}$$

and

$$\text{universal explainer} : \text{universally explaining}.$$

The noun names the capacity. The verb names its continuing and evolving realization.

Universality is therefore not omniscience. It is the absence of a domain-specific ceiling on explanatory creation, given sufficient knowledge, tools, criticism, time, and physical resources.

The finite animal is remarkable precisely because no supernatural exemption is required. A natural organism can explain processes billions of light-years away, build instruments that reveal structures no unaided sense could detect, and use machines to create further explanations. The discovery of the universal explainer is itself a product of nature capable of explaining nature.

4. Physics Reaches the Explainer

The same standard travels farther. Whatever our best physics says about reality applies to people too. At ordinary human scales, you appear simple enough as an object: one body, one location, one continuous self, one reality. Deeper physical explanations need not preserve those ordinary intuitions unchanged.

Biologically, you are composed of cells. While some cells are replaced over time, several types are never replaced. Biological continuity and discontinuity coexist. That is strange when you think about it, but not a problem when you don't.

Similarly, quantum theory describes reality under conditions remote from everyday experience. Serious explanations of quantum reality entail an ontology very different from the classical objects through which ordinary life becomes intelligible.

We do not need to travel to quantum theory here. One consequence is enough. Mutually incompatible theories do not become consilient merely because they are called interpretations. If they make different claims about what exists or how physical reality behaves, then the conflict remains explanatory, not interpretative.

You receive no exemption. Our best physical explanation entails that the everyday object called *you* is only a dynamically stable classical approximation to a deeper physical reality. If so, that consequence belongs to the explanation of personhood too. Explanation does not abolish mystery. It moves it as it travels. The mystery is no longer that people belong to a realm beyond explanation.

Our unresolved problem remains how an evolved physical system creates genuinely new explanatory knowledge. We observe the phenomenon. You can notice a problem no one else noticed and create a conjecture that did not previously exist. You can invent a criterion by which earlier solutions become inadequate. You can criticize an explanation you previously used and thereby change the problem itself.

We can describe many conditions surrounding these events. We can explain neural systems, model learning, describe biological development, and build technologies that retrieve, recombine, transform, and extend enormous bodies of objective knowledge.

But the causal bridge to explanatory creativity remains incomplete. That is an open problem, not evidence of a fixed limit.

$$\text{presently unexplained} \neq \text{intrinsically inexplicable.}$$

Artificial intelligence sits downstream of that gap. Current AI can manipulate objective knowledge at scales inaccessible to unaided human attention. It can retrieve, transform, compare, summarize, generate, calculate, simulate, and operationalize. But capability alone does not settle personhood.

The unresolved questions concern explanatory origination. Does the system create genuinely new problems? Does it create the evaluative structure by which those problems matter? Does it criticize and reorganize its own explanations as a continuing epistemic subject? Those questions remain downstream of the mechanism we do not yet understand.

If explanatory creativity is eventually explained, that explanation would itself become objective knowledge. If the relevant causal structure could then be instantiated in another physical substrate, artificial personhood would become physically possible. AGI would not require magic; it would require a previously unexplained property of people to become understood well enough to instantiate technologically.

The boundary between tool and person would then move. The possibility of that movement does not establish that the boundary has already disappeared.

Now turn the comparison around. Much discussion compares the biological human with the artificial system: which is smarter, which will surpass the other, which tasks belong to which. But people have never realized explanatory universality through biological cognition alone. Language and writing, mathematics and instruments, computers and networks already extend what a biological person can practically explain and do. AI belongs to that lineage of tools.

AI can enlarge the practical region of explanatory space available to a person. It can bring distant fields into one problem situation. It can preserve and transform more explicit structure than biological memory can comfortably retain. It can rapidly generate candidate transformations or expose contradictions across large bodies of objective knowledge.

That does not make the tool the universal explainer. It can make universal explaining more practically realizable for the universal explainer (that is *you*). This is more than a productivity change. It changes the epistemic possibility space in which explanation occurs.

And that creates the decisive turn. Everything so far has described structures that apply to the class of people. You are an instance of that class. But the class does not contain your explanation.

5. The Class Does Not Explain You

Collect the universal result first. A person is:

- an evolved organism;
- an evolving epistemic system;
- a physically instantiated process;
- computationally constrained;
- capable of constructing abstractions;
- capable of creating explanatory knowledge;
- capable of extending that activity through objective knowledge and tools;
- a subjective locus whose inner world remains unavailable to other people directly;
- a fallible entity capable of criticizing inherited explanations and constructing new ones.

These explanations operate at different levels. None replaces the others or possesses jurisdiction over the whole person, yet they must remain mutually possible because they describe one reality. That is already richer than the phrase *universal explainer* alone.

It also creates a new danger. We now possess an explanation of the **class** of persons and can mistake it for an explanation of the instance. You are not the class.

Suppose every universal claim above is correct. We know that you evolved and that your cognition is physically instantiated. We know that you can in principle create explanations, that your mind contains explicit and inexplicit knowledge, and that your history is path-dependent. We also know that you inhabit a world of objective knowledge while your inner experience remains accessible only from your own locus.

We still do not know what problems organize your life, which abstractions have disappeared into familiarity, or which inherited explanations are under tension. We do not know which experiences changed your evaluative criteria, which skills let you notice distinctions that remain invisible to others, or which explanation of yourself you are currently revising.

The explanation of the class therefore enlarges the mystery of the instance. It does not shrink it. The more we understand about people generally, the more causal depth becomes visible in an individual person.

And the instance is stranger still because you participate in your own explanation. A rock does not construct an explanation of itself. A tree does not write an autobiography and then alter its future because the autobiography exposed a contradiction. You can.

You are therefore both an object to be explained and an explainer participating in the construction of the explanation. To understand the breadth is to recursively operate that universal explainer.

That participation gives self-explanation creative latitude without making it arbitrary. A label such as *shy* can lose explanatory force; what seemed like a personal defect can

be reformulated as a badly posed problem. An inherited religion or national identity can be reinterpreted, and the criteria by which vocation or success were once judged can change. Experiences that remained inexplicit can acquire names, and a social category can be reconstructed into a different personal abstraction.

The creativity is real because the explanation can change what happens next. The constraint is real because the world does not change merely because the explanation does. Your body still exists. History still happened. Other people remain independent people. Physical constraints remain physical constraints.

Reality constrains the space in which self-explanation occurs without dictating one permissible self-explanation. That changes the relation between constraint and freedom.

Physics removes impossible transformations and thereby reveals possible ones. Evolution constrains stories about biological origins while revealing the plasticity produced by evolutionary history. Computation constrains physically realizable transformations while showing how the same general machine can perform radically different informational transformations. Epistemology constrains appeals to authority and finality while leaving inherited explanations open to criticism.

Constraint is therefore not merely the opposite of freedom. Better knowledge of a constraint can enlarge the space of possibilities that becomes visible.

Abstract it.

$$\text{better explanation of constraint} \rightarrow \text{larger known possibility space.}$$

You are not free because you are unconstrained. You are clearly constrained. A deeper form of freedom becomes possible because you can create knowledge about constraints and use that knowledge to discover possibilities that were previously invisible.

That result creates a test. If self-explanation is open-ended without being arbitrary, what happens when a personal abstraction and a universal explanation both make claims about the same reality?

6. One Reality, Open-Ended Self-Explanation

Consider a contemporary collision. Humans are a sexually reproducing species with a biologically structured reproductive system. That derives from a biological explanation.

Humans are also explanatory persons capable of creating open-ended abstractions about themselves, their roles, their bodies, their relationships, and their social worlds. That derives from an epistemic explanation.

The two claims need not be rivals. A biological account of sex does not, by itself, contain a theory of gendered identity, role, symbolism, presentation, or self-understanding.

Abstract it.

The inference

biological reproductive structure → only two forms of gendered self-explanation

requires additional explanatory premises hidden in the relation.

Likewise, the creative construction of gendered meanings does not make biological reproduction unreal. The inference

gender is creatively constructed → biological sex is unreal

also hides the relation.

Consilience preserves what is true in both domains while exposing the incompatible overreach. The result is not a political compromise. It is an opportunity for explanatory consilience.

A biologically constrained animal can also be an epistemically open-ended person. Reproductive structure does not specify the number of abstractions people can create about gendered meaning, and open-ended gendered meaning does not rewrite the biological processes that made the organism possible.

The two explanations meet in the same person. Because they meet in you, the case makes the earlier distinction concrete. Your ability to construct explanations of yourself does not imply that anything goes. A self-explanation can conflict with physical reality or another person's independence, reproduce inherited error, become self-sealing, or fail to explain the experience it was created to organize.

Your explanatory freedom is open-ended, but the explanations remain criticizable.

open-ended ≠ arbitrary.

That is where realism and freedom meet. You may create an unbounded set of candidate explanations of yourself; reality does not guarantee that every candidate is equally good.

Relativism can appear attractive here because individual minds really do differ. People occupy different loci, inherit different knowledge, and reconstruct the same reality through different histories. But different perspectives do not entail different realities. There is only one reality.

If one person's truth and another person's truth are insulated whenever they conflict, the contradiction disappears only because criticism has been disabled. That repeats the jurisdictional error at the level of reality itself. One field has its truth and another field another; one identity has its truth and another identity another. The conflict is partitioned rather than explained.

Different perspectives can contain different knowledge while remaining answerable to the same reality. Their explanations may concern different problems or operate at different levels. They may expose errors in one another. But when they make incompatible claims about the same structure, the conflict remains a problem. Relativism

does not solve that problem. It hides it. Returning to one reality does not erase subjective locality.

Universal explanations connect you outward. Evolution places you inside the history of life. Physics places you inside the same reality as every other physical process. Computation places your transformations inside the general space of physical possibility. Objective knowledge connects you to explanations created by people long dead and far away. Technology extends your reach, while society surrounds you with language, norms, artifacts, institutions, and criticism.

Yet your mind remains solitary and local. Other people can communicate with you, model you, know you intimately, and reconstruct parts of your explanations. They do not occupy this locus.

You are therefore both radically connected and radically alone. Your body belongs to the same reality as everything else. Your knowledge is saturated with other people's knowledge. Your tools embody civilization, and your concepts arrive through histories you did not create. Yet experience, problem formation, reconstruction, and self-explanation occur from this locus.

Personhood contains both structures. Neither eliminates the other. The universal explanation has now returned to the singular person without dissolving either reality or individuality. And that changes what can be said about differences among people.

7. From Personhood to Equality and Rights

Differences in present knowledge do not establish differences in epistemic kind. A child may know less than an adult, a patient less medicine than a physician, a passenger less about aviation than a pilot. One person may also be stronger, wealthier, more educated, more articulate, more experienced, or more technically capable than another.

Those differences are real. But if each remains a person in the same explanatory sense, difference does not by itself establish a hierarchy of personhood.

Historically, structures of domination have often depended on explanations that converted difference into rank. Women and children were cast as naturally fit for subordinate or incomplete status. Racial categories were promoted into explanations of intelligence, character, or civilizational capacity. Strength or power could be treated as evidence of a natural entitlement to rule or of a different kind of mind.

The universal account of personhood makes those transitions impossible to sustain. A population difference is not a person. A biological difference is not an explanation. A developmental difference does not establish a lesser explanatory kind. A difference in present capability does not reveal a fixed ceiling on future knowledge.

Anyone who would derive lesser personhood from such differences needs an explanation of the mechanism. The category alone no longer does the work. That changes the epistemic landscape in which rights can be considered.

Rights can then be approached not only as declarations issued by authority but as tentative solutions to recurring problems created by the existence of fallible, corrigible, mutually independent people. Freedom of thought and speech protect problem formation and criticism. Freedom of association enlarges possible cooperation. Bodily protections and equal legal standing resist the conversion of one person into another person's instrument or of social category into ontological rank. These are the signals of explanatory reach.

The question becomes less:

> Which authority granted this person a right?

and more:

> What explanation permits one person to deny another the conditions under which a person can remain a corrigible epistemic agent?

That question is harder because personhood no longer supplies a shortcut from difference to rank. Personhood does not produce a finished political program or settle every moral problem. It changes which explanations remain available.

8. Personhood Remains a Mysterious Explanation

There is an old fear that explanation diminishes its object. If evolution explains us, perhaps we become merely animals; if computation applies to us, merely machines; if physics applies to us, merely matter; if psychology or cognitive science explains experience, perhaps mystery disappears. But explanation need not make the person smaller. It can make the structure deeper.

What disappears is not mystery but the insulation of mystery: *this is beyond explanation.* In its place appears a larger field of wonderful problems.

We can contemplate, *What is a person?*

There is no final explanation; the explanations will continue to travel. Personhood is an explanation constrained by several universal structures. These explanations do not collapse into one another. They describe different aspects and levels of the same reality, so they must remain mutually possible without any one of them claiming the whole person.

A new universal explanation may change the picture dramatically. If it does, personhood changes with it. That is what follows from letting explanations reach you rather than protecting the concept of the human from criticism.

We began by resisting explanations that made you smaller. Again and again, the correction was to return to you, the vast and solitary individual.

But returning to you can't be the final move. Eventually you have to become the problem.

The deepest epistemic problem now resides inside the explainer. You create explanations, recognize problems no one has noticed, and invent criteria by which earlier solutions become inadequate. Yet we still lack an explanation of how you do it.

Does the absence of an explanation tell us anything about whether the process itself can be explained?

That is where the mystery moves and mystery suggests its own next problem.


What Moves You

XVI. Mount Mysterious



There are parts of personhood we are strongly inclined to protect from explanation: creativity, insight, beauty, art, love, grief, consciousness, the soul, and the feeling that another person contains something no mechanism could ever exhaust.

Sometimes the resistance is aesthetic. A scientific explanation seems to flatten the thing it explains. Sometimes it is moral: if love is physical, perhaps it becomes less meaningful. Sometimes it is spiritual: if the soul can be explained, perhaps it was never a soul. Sometimes it is phenomenological. A thought appears. A melody arrives. You suddenly understand. Nothing in conscious experience seems to contain the steps that produced the change.

The tempting conclusion is that something essential must lie beyond explanation. What happens instead when our best explanations reach the parts of you that seem most resistant to them.

The central conjecture is not that mystery is unreal. It is more demanding:

> **What appears as mystery may mark a boundary in our present explicit knowledge rather than a boundary in reality.**

You can fail to explain a process that is nevertheless physically real. A process can remain inaccessible to introspection while remaining causally continuous with the rest of you. A work of art can exceed every explicit criterion we presently possess without

ceasing to arise in reality. Another person's belief can be false while remaining deeply intelligible within the epistemic mind through which that person encounters reality.

The mystery remains, but as a problem rather than an exemption.

1. The Apparent Leap

Creative thought often arrives discontinuously. You struggle with a problem for weeks and suddenly see the solution. You walk away from a bug and wake with an answer. A composer hears a phrase that had never existed before. You carry a moral conviction for years and then, during one conversation, can no longer believe it. The explicit experience has the structure of a jump.

The transition itself may be unavailable to you. That absence is easy to interpret metaphysically. If you cannot consciously reconstruct the steps, perhaps there were no steps. Perhaps creativity arrives from somewhere else, genius is discontinuous by nature, or the soul intervenes.

Abstract it.

Let M_i represent the epistemic mind of person i. Use this abstraction to distinguish one reality from the many individual minds through which reality becomes intelligible:

$$R \neq M_i.$$

You do not directly possess reality. You encounter it through perception, memory, abstraction, language, expectation, inherited objective knowledge, and explanatory history. Much of your M_i is not available as explicit proposition. You can know without being able to say what you know. You can detect that something is wrong without knowing why. You can play a phrase correctly without describing the motor knowledge that makes it possible. You can dislike a room, trust a stranger, or feel that an argument has gone wrong before you can identify the relevant structure.

Your conscious surface is therefore not the whole epistemic mind. It is one interface with it. That changes what the apparent leap can mean. If explicit experience reveals only a thin slice of a much larger evolving structure, a sudden thought may have a long invisible preparation. The leap may mark the point at which a transformation becomes explicit rather than the point at which the transformation begins.

Imagine several structures inside your M_i changing at once. A memory is reconsolidated. A distinction becomes more salient. A skill becomes easier. A social expectation weakens. An analogy becomes available. A familiar explanation fits less well. A new concept changes what counts as relevant. An emotional association shifts. An evaluative criterion changes slightly.

None of these changes need become explicit, and no single change need be dramatic. Their relations can still alter the problem situation.

Abstract it again.
Schematically:

$$\Delta K_a + \Delta K_b + \Delta K_c + \Delta K_d \rightarrow \Delta M_i.$$

The notation is not meant literally as addition. The structures interact. A small change in one region may matter only because another region has also changed. A new concept may remain inert until an old memory becomes newly salient. A criticism may fail repeatedly and then become devastating because your evaluative criteria have moved.

Your epistemic mind is not a single structure that changes all at once. It contains memories, habits, skills, fears, values, interests, social knowledge, bodily states, unfinished problems, explicit theories, inexplicit expectations, and changing criteria. These structures can conflict. They can change at different rates and become conscious at different times.

You are therefore an epistemic multiplicity without becoming multiple people. The subjective locus remains one. The epistemic structure beneath it can be extraordinarily heterogeneous. One part of you may remain stable while another changes quickly. A new explanation may enter one domain while old habits preserve another. A moral criterion may change before behavior does. A skill may change behavior before language catches up. A new relationship may reorganize salience long before you notice that your goals have changed.

Eventually some consequence of that changing structure crosses the narrow explicit surface. You say, "I just realized something." The realization may be sudden even when the change that made it possible was not.

A simple physical analogy establishes only the structural possibility. Water can be heated gradually. At the molecular level, innumerable local interactions change and energy accumulates. At the level relevant to ordinary experience, a qualitatively different state then becomes obvious: the water boils.

A creative insight is not being claimed to be a thermodynamic phase transition. The point is narrower:

> **Continuous lower-level change can become visible as qualitative discontinuity at another explanatory level.**

The conscious report "now I understand" therefore does not establish that nothing relevant changed before now. The absence of conscious access is not evidence of causal absence.

2. Causal Continuity and the Open-Ended Self

Physics constrains the mystery here. Whatever happens when you think, remember, love, grieve, invent, compose, criticize, or change your mind happens in a physically instantiated system.

That claim does not identify the relevant mechanism. It does not tell us which neural structures matter. It does not explain subjective experience or explanatory creativity. It only prevents those phenomena from being treated as though they occur outside reality. Your body is not a vessel carrying an epistemic process from somewhere else. The epistemic process is physically instantiated.

The unexplained event may be extraordinary, but it cannot be causally arbitrary. Physics therefore constrains what an explanation of you can require. We need not settle every philosophical dispute about determinism to see the general point. Physical states have causal histories. Interactions occur under physical constraints. A thought does not need to appear from nowhere in order to be novel, and a new explanation does not need a breach in locality in order to be creative.

Randomness does not solve the problem. Suppose an unexplained neural event were genuinely random. That might explain unpredictability. It would not explain why the event instantiated a good explanation of orbital mechanics, a new proof, or a melody that transformed music.

Randomness is not creativity. The problem is therefore not how novelty escapes causation, but this:

> **How can causally continuous physical processes instantiate epistemic novelty?**

That is a harder problem because it preserves both sides of the phenomenon: novelty is real, and so is causal continuity.

The same constraint reaches self-transformation. Earlier essays rejected fixed personality as destiny. You can criticize inherited explanations, create new problems, and become radically different. But open-ended transformation does not imply instant revolutions.

You have a history. Your body, skills, relationships, concepts, fears, attractions, and inexplicit knowledge all have histories. You can create a new explicit explanation in an afternoon and spend years reconstructing the rest of M_i around it. Freedom inside a causally continuous system therefore has a particular shape. The future is open, but it is not causally isolated.

3. When Incremental Change Becomes Revolutionary

Problems contain evaluative structure. You do not merely ask a question; you possess some sense, often inexplicit, of what would count as improvement. The difficulty is that those criteria can themselves change.

Objective knowledge records such changes everywhere. A scientific explanation can change what later scientists count as an adequate explanation. A moral innovation can make a previously ordinary practice intolerable. A new technology can alter what users later expect from every technology in its class. A work of art can change what later listeners can hear.

Abstractly:

$$C_t \neq C_{t+1}.$$

The historical change is visible. The mechanism is not. We can say that evaluative criteria change without pretending to know how you create the change. That distinction preserves the problem rather than filling it with a label.

Now suppose several dimensions of your M_i change piecemeal. Your skill improves. Your attention becomes trained. A new abstraction becomes available. A previously invisible contradiction becomes irritating. An old criterion weakens while a new one begins to form.

None of these changes alone looks revolutionary. Then one explicit product appears: a theory, a painting, a moral argument, a line of code, a political idea, a musical phrase. The artifact may reorganize the objective knowledge around it. Its explanatory content may be revolutionary even if its causal history was not.

This gives incrementalism a precise meaning. It is not the doctrine that important change must be small. It is the recognition that complex adaptive systems have histories and dependencies, so large visible transformations can emerge from many smaller changes that never became explicit.

The same architecture helps explain why the parts of personhood we most readily call mysterious often share a feature. Artistic taste, love, grief, beauty, intuition, spiritual experience, moral revulsion, creative inspiration, a sense of meaning, and the soul can all be experienced with extraordinary force. The structures that produced the experience may remain largely inaccessible.

The asymmetry matters. The explicit experience can be vivid precisely where the explanatory mechanism is obscure. The obscurity then becomes part of the phenomenon, and the experience can feel as though it arrived from beyond explanation. But the process may instead lie mostly within inaccessible regions of M_i. The mystery may reflect the architecture of access.

4. Mystery as an Epistemic Relation

There is an important difference between saying "I cannot explain this" and saying "this cannot be explained." The first locates a deficiency in present knowledge. The second makes a universal claim about reality.

That second claim has much greater reach. If some physical process is intrinsically inexplicable, then a further problem appears: why does the same reality that permits the process not permit explanatory access to it? Without an answer, the mystery has only been renamed.

A forever-mystery therefore requires more explanatory work than a temporary one. "We do not know yet" marks the present boundary. "No one can ever know" makes a stronger claim about reality itself.

You are physical. Your art is produced physically. Your grief is physically instantiated. Spiritual experience occurs in a physical person. Your memory is physical. Your changing evaluative criteria are physically instantiated. Your explanatory creativity, however mysterious, occurs in physical reality. Every successful explanation we possess is also created from within that same reality.

None of this guarantees that a problem will be solved. Time, resources, access, history, or civilization may impose limits. But present ignorance does not establish an in-principle boundary.

The resulting epistemic position is modest but strong:

> **No physical phenomenon earns permanent exemption from explanation merely because our present explanations do not reach it.**

There are then two importantly different uses of mystery. One says: there is a real phenomenon here that our present explanations do not adequately explain. The other says: because we cannot presently explain it, explanation must stop here.

Call the first **good mystery** and the second **bad mystery**. Good mystery creates a problem. Bad mystery closes one. Mystery becomes anti-explanatory when a description of ignorance is converted into a prohibition on inquiry, so that the unknown becomes protected precisely because it is unknown.

Several mysteries of personhood remain genuine known unknowns. We do not explain how a physical system creates a new explanatory problem or how evaluative criteria change inside one person. We do not explain why one criticism becomes transformative while another remains inert, or how subjective experience relates to physical implementation. We do not explain why some artistic perturbations reorganize whole traditions, or why an idea can arrive explicitly after years of invisible preparation.

These are not failures to be hidden by a more impressive vocabulary. They mark the present edge of the explanation. Their epistemic role is to remain exposed as problems.

5. Art as a Natural Experiment in Mystery

Other minds provide an unusual source of criticism because human cognitive machinery varies far more than ordinary conversation reveals. Some people visualize vividly; others do not visualize at all. Some possess rich inner speech; others report little or none. Some can re-create music internally; others know music without consciously hearing it. Different people can reach similar outward competence through very different subjective routes.

These differences function as natural experiments. They let us separate a phenomenon from one candidate explanation of the phenomenon. If two people can produce the same kind of objective result without sharing a conscious mechanism we assumed was necessary, that mechanism cannot be decisive. The deeper process remains unexplained, but one explanation has been removed.

Consider two musicians. One possesses extraordinarily vivid auditory imagery. They can hear orchestration internally, rehearse candidate phrases without making a sound, and compare variations inside conscious experience. The other has little or no voluntary auditory imagery. There is no internal concert hall. A melody may be known without being consciously heard. A harmonic possibility may be represented symbolically, relationally, spatially, or in some other form the musician cannot fully describe.

The second musician may need to externalize more of the critical process. They play the phrase, hear it physically, revise it, and play again. The external world supplies a critical surface that the first musician can partially simulate internally.

Their conscious workflows differ radically. Their music need not. The hyperphant and the aphant can therefore travel radically different internal routes toward the same objective musical world.

Once the work becomes objective knowledge, you may have no access to the cognitive route that created it. Two songs can be equally moving, two compositions equally complex, and two performances equally original. You may still be unable to infer which creator heard the piece vividly before playing it and which had to externalize almost every tentative solution.

Abstract it:

$$\text{different subjective machinery} \rightarrow \text{similar objective artistic result.}$$

At least some conscious mechanisms associated with creativity therefore cannot be necessary conditions of artistic quality. The comparison does not dissolve the mystery; it moves it. The remaining question is what process actually carries the creative work.

6. Artistic Soul and Changing Criteria

People sometimes say, "this music has soul." The phrase can be meaningful without yet being explanatory. The work may feel emotionally alive. It may violate a convention in a way that feels necessary or carry traces of struggle. It may surprise while remaining coherent, connect distant parts of your own M_i, change what you are able to hear, or resist every decomposition you presently have. Those are different possible problems. The word *soul* can compress them into one apparent solution.

Once they are separated, the explanatory question changes. Where is the soul supposed to be? It might be located in the artist's feelings, the method of production, the artifact, you as the listener, the relation between the artifact and your evaluative criteria, or the history of the genre. The more precisely the possibilities are distinguished, the less the mystery behaves like one thing.

The problem becomes sharpest when art changes the criteria by which later art is judged. Most art is evaluated within criteria that already exist. A musician can write a competent song, a painter produce a skillful image, or a novelist satisfy the conventions of a genre. Sometimes, however, a work changes the criteria. Listeners begin to hear differently. Later musicians inherit distinctions that did not previously exist. An innovation that first sounded wrong becomes indispensable.

The future criterion is not fully available to the present evaluator. The work may help create the standard by which it is later recognized. That makes genius difficult to identify from inside an existing evaluative landscape without making it mystical. The unknown remains elsewhere: we still do not explain how you create the perturbation that changes the criteria.

Artificial intelligence sharpens that unknown. A system with access to enormous objective knowledge can reproduce, recombine, and optimize against explicit musical structures at extraordinary scale. As evaluative criteria become explicit enough to be represented in objective knowledge, artificial systems should become increasingly capable of satisfying them. The relevant criteria include harmony, rhythm, instrumentation, genre conventions, production patterns, formal structure, and stylistic resemblance.

Listeners may call much of the result derivative or slop, but human beings also produce enormous quantities of derivative work. Mediocrity therefore cannot by itself distinguish the human from the machine.

The more interesting boundary lies in the creation of new problems and criteria. The strongest present human distinction available to this argument is not that every person creates great art; most do not. It is that a person may create a problem or criterion that did not previously exist in objective knowledge.

A tiny perturbation may matter: one strange chord, one unexpected pause, one lyric that changes the interpretation of the whole form, one aesthetic decision that initially

looks like an error. Many minds may encounter the perturbation, and something inexplicit changes in each. The new criterion can therefore become distributed across many M_i before anyone can state it explicitly: listeners begin to hear and judge differently without yet being able to say what has changed. Only later do creators reconstruct the perturbation until the new criterion becomes part of the explicit landscape.

That is miraculous.

The mystery is not that the original artist violated rules. Machines can violate rules. The mystery is how you create a variation that another person experiences as the beginning of a better rule. That returns us to dynamic evaluative criteria.

Not all artistic value works this way. A parent may cherish a child's drawing more than a technically accomplished painting because the drawing belongs to a particular developing person. It records effort, carries a relationship, belongs to a moment in a shared history, and may reveal a new capacity. The artifact is inseparable from the parent's model of the child.

The emotional value is real, but it differs from the claim that the work changed the objective history of art. Value can depend on the relationship between artifact and M_i without becoming a universal artistic criterion.

One can avoid the difficulty by defining art as "the expressive process of a human person." Then the child's drawing is art by definition and AI output is not. The distinction becomes easy because the explanation has disappeared into the category. The process is artistic because a human performed it; the human is special because the process is artistic. Nothing has been explained about why one artifact changes evaluative criteria while another does not.

The same definition excludes any future non-human person before the relevant capacities are examined. Biological membership carries the difference by classification. If personhood is used that way in aesthetics, the categorical error returns: a classification has become an explanatory substitute.

7. From Artistic Soul to Human Soul

The artistic soul points naturally toward the human soul because both name an invisible something that seems to explain why the visible object exceeds the explicit account.

A musician can be technically perfect and said to lack soul. A cruel person can be called soulless. A loved person can be imagined as possessing a soul that survives the body. The soul can be treated as the essence of identity or as what animates the body. It can be treated as what experiences reward or punishment after death, creates art, or grounds moral worth. It can remain invisible and also be declared beyond explanation.

A single word is therefore carrying several different explanatory problems. The pattern becomes clearer when the word is used as an answer:

> Why is the living person different from the corpse? Soul. Why does this music feel alive? Soul. What persists after death? Soul. What gives the person moral worth? Soul. What distinguishes the person from a machine? Soul.

Yet when the causal structure is requested, the answer may become "the soul is beyond explanation." The concept is then asked to perform explanatory work. At the same time, it is protected from the conditions under which explanations can be criticized.

The problem is structural rather than spiritual. If the soul explains, it must have consequences. If it has consequences, those consequences enter reality. Suppose the soul animates the body, preserves memory, influences action, creates artistic expression, or survives death. Each statement asserts a dependency: something happens because the soul exists or acts. Those dependencies must coexist with our other explanations of physical bodies, memory, cognition, death, and causation.

Taking the soul seriously as an explanation therefore does not require banning it. It requires following what the explanation says happens. An unfamiliar domain can remain unfamiliar without acquiring immunity from criticism merely because it is important.

8. Mystery Inside an Epistemic World

The inquiry now turns outward again, toward other people. Mystery reveals something about how you inhabit M_i: you do not experience your own explanations as diagrams sitting beside reality. You experience the world through them.

A child is not experienced as an abstract bundle of propositions. The child is loved. A deceased partner is not experienced as a missing variable; the absence can reorganize the whole experienced world. A religious cosmology is not necessarily stored as a detached list of claims; it can structure family, morality, death, hope, fear, identity, and future expectation.

A mysterious explanation can therefore occupy an enormous region of M_i even when its explicit sentence is small. Its epistemic roots may be vast.

Consider a person who says, "I refuse to believe that the people I love are really gone, that I will never see them again." The sentence can be criticized as a proposition about reality. It is also a doorway. Behind it may sit decades of memory, attachment, shared plans, habits of expectation, family stories, and religious knowledge. It may also contain fear of finality, a sense of justice, the inability to reconstruct identity without the loved person, and the persistent feeling that the dead remain present.

None of this makes the proposition true. It makes the person's problem situation much larger than the sentence. A reply that consists only of "you are wrong" can

therefore miss what was exposed. You may not have presented an argument; you may have exposed a fracture in M_i.

The same structure appears in other beliefs. Suppose a person sincerely believes in eternal punishment. That belief may be integrated with love. The parent says, "I do not want my child to go to hell." Inside the parent's M_i, this can be a profound protective problem. The imagined consequence is larger than physical death, so the urgency can exceed the urgency attached to ordinary danger.

Once that structure is visible, coercive behavior can become explicable without becoming correct. The parent may restrict speech, friendships, sexuality, books, technology, or belief and experience each restriction as care. The problem is not necessarily an absence of knowledge. A large body of knowledge may instead be organized around a false explanation of R.

That distinction becomes important when criticism reaches the person carrying the belief.

9. Criticism, Reconstruction, and Coercion

A criticism does not simply strike a false proposition with reality. It's creative. You reconstruct it, and the reconstruction changes your problem situation.

That structure creates a possibility that is easy to miss: criticism itself can become coercive.

Suppose someone reveals a fragile explanation that is presently holding together a devastated epistemic mind. The critic responds, "that belief is false; abandon it." The critic may possess the better explanation of R. Even so, the intervention can function as:

$$\text{deep problem structure} \rightarrow \text{forced epistemic replacement.}$$

The destructiveness does not make the criticism false. The falsity of the belief also does not establish that forced replacement can complete the recipient's reconstruction. The relevant distinction is:

$$\text{criticism} \neq \text{compelled reconstruction.}$$

A criticism can expose a problem. It cannot perform another person's reconstruction for them.

Pain does not define the difference. Some valuable criticisms are painful. An explanation can collapse. A cherished identity can become untenable. You can discover that you harmed someone. A scientific theory can fail. A relationship can end. Error elimination is not obligated to preserve comfort.

The coercive element appears when another person attempts to close your epistemic process by requiring a successor explanation: "you must now adopt this explanation."

The space to reconstruct, reject, defer, or modify is then removed. So is the space to create a successor problem. Creative criticism has become authority using the vocabulary of truth.

Two parents make the structure easier to see.

The first sees a child step toward an oncoming car. Within the parent's M_i:

$$\text{moving car} \rightarrow \text{possible catastrophic bodily harm} \rightarrow \text{urgent intervention.}$$

The parent grabs the child. The intervention is commonly treated as justified.

The second parent believes that some action could condemn the child eternally. Within that parent's M_i:

$$\text{forbidden act} \rightarrow \text{possible eternal suffering} \rightarrow \text{urgent intervention.}$$

The parent constrains the child. That intervention may be condemned.

Yet the abstract problem structures can be strikingly similar. Both parents believe they are preventing catastrophic harm. The difference is not that one cares and the other does not. It lies in how the relevant structures in each M_i correspond to R.

An action can therefore be intelligible within a person's epistemic mind without the explanation producing it being true:

$$\text{rationality within } M_i \neq \text{truth about } R.$$

The parent who fears the car may possess a good explanation. The parent who fears hell may possess a false one. Understanding either action requires identifying the problem the parent believes they are solving.

That is not relativism. There are not two realities. There are two fallible epistemic minds, both encountering one R, both open to criticism, and neither accessing R except through its own evolving M_i.

10. The Roads Between Epistemic Minds

This changes the interpersonal problem. Suppose I believe your conclusion is false. I can criticize the conclusion directly, and sometimes that is enough. But if the conclusion is integrated across a large region of your M_i, my criticism may never reach the structure sustaining it. It may be heard as noise, threat, evidence that outsiders cannot be trusted, or proof that the sacred thing is under attack.

The disagreement then persists for reasons that cannot be reduced to missing information. You may possess an enormous amount of knowledge. The criticism may fail because it must travel through an epistemic landscape I do not understand.

Understanding another person therefore requires more than locating an explicit belief. It requires reconstructing the road by which that belief became part of reality-as-experienced.

This also explains why more argument does not necessarily produce less disagreement. A criticism enters M_i and creates a problem. You generate new explanations. Those explanations protect or reorganize the existing structure. Your epistemic mind becomes richer while the disagreement becomes more sophisticated.

So:

$$\text{more argument} \nRightarrow \text{less disagreement.}$$

Sometimes:

$$\text{criticism} \rightarrow P \rightarrow TS \rightarrow \text{stronger conflicting knowledge.}$$

The result is not an absence of knowledge. It is knowledge against knowledge. Deep conflicts can therefore become more durable among intelligent and highly informed people.

Private mystery can also become social. You externalize an explanation. It enters language and becomes story, doctrine, ritual, music, institution, law, or tradition. Later people can reconstruct it without having your original experience.

The mystery has become objective knowledge. It can preserve insight or error, create beauty or authority, help a person survive grief, or justify coercion. The same externalization that makes criticism possible also allows an explanation to persist beyond the individual who first created it.

When mysterious explanations become socially authoritative, their causal reach changes. A private belief can become doctrine; doctrine can become role; role can become institution; institution can become coercive structure. The explanation now acts through objective knowledge.

The social system does not possess a mind. Individuals reconstruct the explanation. They teach it, enforce it, fear its consequences, and create new explanations around it. Mystery can therefore become socially durable without ceasing to depend on solitary minds. Control need not arise from cynicism or power seeking. It can arise from sincere explanatory commitment, which can make the structure harder to criticize rather than easier.

You can sometimes travel toward another person's epistemic mind, not by entering or merging, but by conjecture. You build an inexplicit model, ask what problem the other person is solving, let them criticize your reconstruction, and discover where your own categories fail.

Mystery gives that practice a harder purpose. Epistemic travel can be necessary for effective criticism. If you do not understand the structure in which your explanation

lives, you may attack a sentence while leaving the explanation untouched. You may even destroy a fragile scaffold without understanding what else it supports.

Taking another person's explanation seriously and personally therefore begins by allowing it to become a problem for my model of them. But understanding is not endorsement.

Abstract it.

To reconstruct M_j is not to adopt M_j. My model of your epistemic mind remains:

$$\widehat{M_j}.$$

It is a conjecture, and you can criticize it. Even if I reconstruct you accurately, your explanation remains criticizable against R.

Two questions must therefore remain separate:

> Do I understand what you mean and what problem your explanation solves?

and

> Is the explanation true?

The first is interpersonal reconstruction. The second is criticism. Confusing them produces two symmetrical errors: because I understand you, you must be right; or because I think you are wrong, there is nothing to understand. Both stop inquiry too early.

The critic has no privileged escape. I do not criticize you from outside all epistemic worlds; I criticize from mine. My explanations may correspond better to R and survive criticism yours does not, but I still reconstruct reality through my own abstractions, language, history, and inexplicit knowledge. That does not make truth relative. It makes fallibility universal.

The critic can therefore become part of the problem. You can mistake confidence for direct access. You can confuse possession of a better explanation with possession of your reconstruction process. You can treat criticism as correction rather than creation, and become coercive while believing you are merely supplying truth.

11. Consilience Without Reduction

The inquiry returns to fear. If art is physical, does art become less? If grief is physical, does grief become less? If spiritual experience is physically instantiated, does meaning disappear? If creativity is causally continuous, does creativity become mechanical?

Those conclusions do not follow from physical instantiation. A lower-level explanation does not erase the higher-level phenomenon. The art remains. The grief remains. The insight remains. *You* remain.

Consilience requires the explanations to coexist. It does not require one vocabulary to replace all the others. Preserving every real phenomenon while refusing explanatory exemptions is harder than reducing one level to another, and harder than leaving the mystery untouched.

The choice between explanation and awe is false. Mystery is not preserved by ignorance. A better explanation can reveal that the phenomenon was stranger, deeper, and more connected than the mysterious prohibition allowed us to see.

At the same time, objective knowledge shows that evaluative criteria really do change. Art changes. Science changes. Morality changes. You may create a perturbation that later minds inherit as a new way of judging. We can observe that transformation without yet understanding its origin or transmission.

Rejecting explanatory immunity does not make the people who carry mysterious explanations unintelligible. Their explanations live somewhere. A belief about death may sit inside love, memory, identity, fear, family, religion, and the inability to reconstruct a world after loss. A belief about hell may create a sincere protective problem for a parent. Understanding that problem does not make hell real; it makes the parent intelligible.

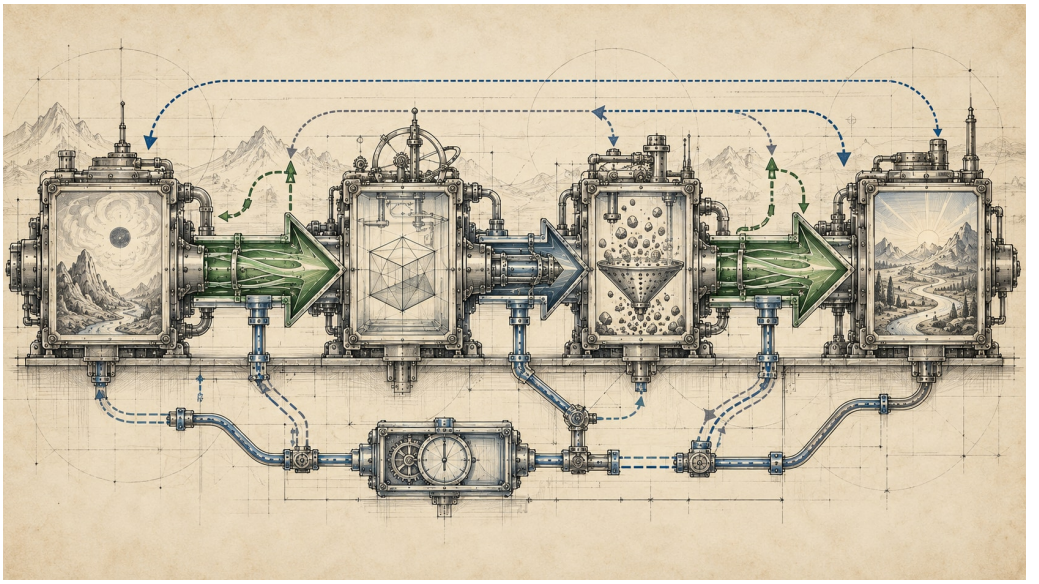
Exploring another person's mystery therefore requires travelling the road available to that person. The problem they are solving has to be reconstructed, and that reconstruction remains open to their criticism. Understanding must remain distinct from agreement, while every road remains oriented toward one reality.

You can reconstruct another person's explanation without accepting it. You can expose an error without compelling the person to reconstruct the knowledge that sustains it. And even when a criticism is understood, it may fail to change the problem.

Why does one criticism remain inert while another reorganizes an epistemic mind? What makes a person move from one explanatory state to another?

We have become increasingly precise about the problems, explanations, and criticisms through which knowledge grows. But something essential remains hidden in the movement between them.

XVII. The Arrows of Knowledge



One schematic has endured:

$$P_1 \to TS \to EE \to P_2.$$

A problem. A tentative solution. Error elimination. A successor problem.

The cycle is useful because it is compressed. It can describe a scientific discovery, a child's learning, an argument failing under criticism, or the reconstruction of an explanation inside one solitary mind.

But compression hides things. Our attention repeatedly falls on the states. What is a problem? What makes a tentative solution explanatory? What does criticism do? How does a successor problem differ from the problem that came before it?

Now return to the schematic. What are the most neglected parts? The arrows.

What moves you from one explanatory state to another? Why does a problem suddenly feel sufficiently formed that a tentative solution becomes possible? Why does a conjecture that felt promising an hour ago suddenly demand criticism? Why does one criticism reorganize the problem while another disappears? Why do you abandon one problem and become fascinated by another? Why does a solution arrive during a walk? Why can you stare at the same problem for hours and feel no movement at all?

We do not know. The arrows are compressed too. They are not proposed here as neural mechanisms, cognitive modules, neurotransmitter flows, or literal structures

inside the brain. They are epistemic abstractions. They make visible something that has been present throughout but remained almost typographically invisible:

> **You do not merely inhabit explanatory states. You move among them in mysterious ways.**

That movement may be as important as the states themselves. And we may have built psychological, educational, institutional, and moral vocabularies around the states while neglecting the arrows.

1. The Arrows Are Not a Clock

The obvious mistake would be to read the epistemic cycle as a temporal schedule or linear sequence.

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2$$

can look like:

1. finish defining the problem;
2. generate a solution;
3. criticize the solution;
4. formulate the next problem;
5. repeat.

No one lives like that.

A scientist can criticize a tentative solution before the original problem is fully articulate. A programmer can abandon a bug, return two days later, and discover that the original problem has changed. You can carry incompatible explanations for years before some event makes their conflict explicit, or wake with a candidate solution to a problem you had stopped consciously thinking about.

So the arrows are not instructions. They do not say:

> now proceed from state A to state B.

The diagram is a map of epistemic relations. It identifies recognizable conditions in the growth of knowledge. Your lived process can move forward, backward, sideways, recursively, intermittently, and sometimes apparently nowhere at all.

You do not owe the diagram linearity. The diagram provides explanatory usefulness.

This distinction matters because states are easier to see than transitions. A problem can be written down, an explanation modeled, a criticism recorded, and a successor problem articulated. Objective knowledge favors these states.

The arrows are harder to capture. What moves you from *I don't understand* to *maybe this explains it*, or from *this explanation seems good* to *something here is wrong*?

You may not know. The observer may see only a jump. The biological machinery remains hidden. The epistemic movement may become visible only after it has occurred.

That asymmetry creates a bias. States look like the serious intellectual content because they can be inspected after they have been externalized; transitions look incidental because much of what produced them remains hidden. Call this the **state-transition bias**.

Imagine, only as an analogy, a computational architecture in which one agent processed the current state while another orchestrated movement among states. The first might solve equations, inspect evidence, compare alternatives, or refine an explicit explanation. The second might determine that the problem has changed, that the present solution should be criticized, that a different problem deserves attention, or that the system should stop processing entirely and wait.

We would not say:

> the first agent is intelligent; the second is merely distracted.

The architecture would depend upon both. There is no claim that your brain contains a state processor and an orchestration processor. The point is conceptual. We readily value processing once the problem frame is stable. We have much less vocabulary for valuing the movement that changes the frame.

2. Your Arrows Are Harder to See

The state-transition bias becomes socially important because your own transitions and another person's transitions are not equally legible to you.

Suppose someone receives a well-defined problem and processes it rapidly. They calculate quickly, remember relevant facts, write clearly, and produce many correct transformations without leaving the problem frame. We call them intelligent.

Now suppose another person moves rapidly across states. A problem suggests a conjecture. The conjecture reveals a contradiction. The contradiction makes another problem salient. That problem connects to something apparently distant. The person leaves the original subject, returns later, and produces an explanation whose path nobody else followed. We call them scattered.

The difference is not necessarily that one mind is doing more epistemic work. One person's work remains legible inside a state the observer already understands. The other repeatedly destroys the observer's anchor.

Externally, the observer keeps having to ask:

> Wait. What are we talking about now?

Internally, the transitions may feel obvious.

That difference matters because social judgments are made from outside. You inhabit the epistemic mind that makes your own transitions intelligible. Another person sees only the surface: a change of subject, a criticism, a pause, a sudden enthusiasm.

They have to reconstruct the road. Sometimes they can. Sometimes they cannot. You may then appear less coherent than you are. Or more coherent than you are. The visible transition is underdetermined by the invisible epistemic landscape that produced it.

Inner speech gives one unusually accessible example. Not everyone reports a continuous inner monologue, so inner speech cannot be treated as universal machinery of thought. But where it is present, its structure is revealing.

It rarely sounds like a clean essay. It may sound more like:

> I need to finish this. Why am I avoiding it? Actually, the problem is not the work. It is that I don't believe the premise. Wait, that is the same issue from yesterday. No. Maybe the issue is that I am solving the wrong problem. I should write this down. I'm hungry. That reminds me—

The words are not merely processing one stable state. They are traces of transitions. Problems are reformulated. Candidate explanations appear. Criticism interrupts them. Memories become relevant. Evaluative criteria change. Attention moves.

For you, where inner speech is present, the monologue can be arrow-laden. Where it is absent, the same movement may appear differently: a shift in bodily orientation, a sudden certainty that the present direction is wrong, or a new explicit conclusion appearing without a narrated path.

The implementation can differ. The arrow remains a deliberately abstract description of movement.

3. The Epistemic Mind Moves

The metaphor of an epistemic landscape is useful. Your epistemic mind contains explicit explanations, inexplicit knowledge, memories, skills, interests, abstractions, values, expectations, unfinished problems, and inherited objective knowledge.

But landscapes sound static. The arrows change that. Your epistemic mind is not merely a place through which you travel. The travel changes the place. A criticism can make an old path unusable. A new explanation can connect regions that previously seemed unrelated. A skill can make new problems visible. A failed project can reorganize years of prior experience.

The difference need not come from acquiring a new explicit proposition. Movement itself can reshape the problem field. Perhaps the epistemic landscape is better imagined as **landscaping**: not a finished terrain, but an evolving one. You move through it while

continuously altering the conditions of later movement. That is path dependence made dynamic.

The same shift makes an old question sharper.

Earlier we asked:

> where does the problem come from?

Here we ask:

> why this arrow? why now?

Why does one problem release its grip while another suddenly becomes alive? Why does a criticism that was available yesterday become devastating today? Why do you finally feel ready to stop criticizing and construct again? Why does an irrelevant memory become relevant? Why can boredom suddenly become curiosity?

We do not possess an explanation. And the notation should not be mistaken for one.

To write:

$$P \rightarrow TS$$

does not explain the causal machinery by which a problem becomes a conjecture. It makes the transition explicit enough to become a problem.

That is already useful. The unknown becomes more visible.

Sometimes what becomes intelligible first is not the destination but the need to move.

You may think:

> I understand the problem well enough. I need to start constructing something.

Or:

> I have spent long enough decorating this explanation. It is failing. I need to criticize it.

Or:

> I know the solution failed, but I do not yet know what the successor problem is.

Or simply:

> I cannot productively inhabit this problem right now.

The state is not resolved in the sense of closure. But direction has become intelligible.

You may still be confused. You may have no solution. You may not even know what problem comes next. Yet something fits:

> **I need to move.**

The arrow can become consilient before the destination does.

4. When Direction Feels Alive

That brings us to a deliberately unserious-sounding meme.

Fun.

Modern intellectual culture does not usually grant fun much epistemic dignity. Fun belongs to leisure. Serious work belongs to effort. A difficult activity is respectable because it is difficult. A pleasurable activity is suspect because pleasure seems to explain why you are doing it.

We inherit slogans such as:

> no pain, no gain.

The slogan contains a truth in some problem situations. Bodies adapt to load. Skills often require repetition. Difficult knowledge cannot always be acquired without discomfort.

But the meme can quietly acquire a much broader moral meaning:

$$\text{pain} \rightarrow \text{seriousness}$$

and:

$$\text{fun} \rightarrow \text{indulgence}.$$

That deserves criticism. Suppose fun is not treated as pleasure alone. Suppose we use it provisionally to name the experience of an epistemic direction becoming alive. A child follows a question because it is fun. A scientist cannot stop thinking about an anomaly. A conversation runs for hours because each criticism creates another problem.

The common element is not ease. These activities can be difficult, exhausting, frustrating, even painful. You continue because the next arrow feels available. Call this provisional signal the **fun criterion**.

So the conjecture is not:

> fun is truth.

Nor:

> fun is always good.

Nor:

> you should only do what is fun.

It is narrower:

> **Fun may sometimes be the conscious surface of a much larger inexplicit epistemic process drawing you toward a transition.**

It is not a utility function. You are not necessarily calculating which option will maximize fun. The relevant option may not have existed a moment earlier because the problem, the criteria, or the landscape may itself be changing. Fun may be one way the movement becomes experientially visible.

Childhood makes the possibility hard to ignore. A small child moves constantly among problems: touch this, ask why, stack the blocks, lose interest, return, invent a game, change the rules, become absorbed in something the adult did not notice.

We usually interpret much of this generously. The child is learning, exploring, developing, playing. We do not require the child to explain the epistemic value of every transition. We assume that an enormous amount of development is occurring beneath explicit articulation.

Then something changes. At some poorly explained point, the cultural model shifts. The same kind of movement becomes distraction. Remain seated. Follow the assigned problem. Finish before switching. Do not follow the unrelated question. Fun becomes a reward after work rather than a possible feature of knowledge creation.

The epistemology has not obviously changed. The child remains a problem-generating, conjecture-forming, criticism-sensitive person. So what explanation licenses the boundary? Why should self-directed epistemic movement count as development at six and indiscipline at sixteen?

The problem is narrow. Externally ordered learning can become so dominant that your own arrows are interpreted only as interference, or evidence of distraction. That interpretation requires more explanation than it usually receives.

5. When Attention and Direction Diverge

The screen-time problem now looks different.

The conventional observation is easy:

> the child has spent too long on the screen.

The adult sees duration, stillness, perhaps inactivity.

But what does the child experience? Perhaps nothing worthwhile. Perhaps compulsive repetition or escape. Perhaps a social or technical problem. Perhaps a creative project or a game containing nested problems and rapidly changing criteria. Perhaps several of these at once.

Removing the screen may abruptly terminate a dense epistemic landscape whose arrows the adult cannot see.

The earlier error was to infer psychological content from the medium. The arrow adds another caution:

> **The observer may see a static screen while the child experiences rapid epistemic movement.**

Physical stillness can contain cognitive travel. Visible activity can contain epistemic stasis. The external form does not settle the explanatory question.

The same issue appears in bureaucracy, repetitive workflow, and the classroom exercise whose answer everyone already knows. A process may specify the problem, the acceptable solution, the sequence, the language of failure, and the next task.

The arrow makes the mechanism of epistemic confinement more tangible.

The process does not merely assign a state. It assigns the transition:

$$P_A \to TS_A \to EE_A \to P_B.$$

You are expected to remain on the prescribed sequence. When the required transition and your live epistemic direction diverge, the process creates what we can call a **forced arrow**.

But your epistemic mind may have moved. You may see that P_A is badly formulated. You may have generated TS_C. You may have encountered criticism the process does not permit. You may already be at P_D. Or no arrow may presently exist.

The institution nevertheless demands:

> continue.

This can create a distinctive form of disorientation. You experience one epistemic direction while being required to enact another. The conflict may feel like boredom, restlessness, confusion, irritation, or even shame. The authority may then interpret the symptoms as defects in you.

Why can't you focus? Why can't you just finish? Why do you keep changing the subject? Why are you making this complicated? Why are you so unmotivated?

You may not know. Your own arrows are not necessarily available to introspection. You simply know that remaining here feels increasingly difficult.

Creative blockage gives the inverse case. You have a problem. You want the solution. You sit down. Nothing comes. So you push harder. Think harder. Concentrate. Force yourself to remain. But the next epistemic transition is not available. The problem remains a problem. No conjecture feels alive. No criticism opens anything. No successor problem becomes visible.

It can feel like pushing a rope. The effort acts upon the present state without producing the transition. Then you stop. You walk, talk to someone, sleep, or do something apparently irrelevant. Later the arrow appears.

The image should not be romanticized. A walk is not a magical creativity technology. Sometimes the problem remains stuck. Sometimes disciplined persistence is exactly what produces the breakthrough.

The narrower point is that explicit effort does not seem to control every transition. Your inexplicit epistemic mind continues changing outside the narrow surface of conscious problem-solving. When the arrow finally appears, you may experience it as sudden.

We should not infer causal discontinuity from conscious suddenness. Perhaps the conscious arrival of the arrow is itself one of the things you experience as creative movement.

Attention makes the stakes higher. Difficulty sustaining attention can cause real suffering. It can damage education, work, relationships, safety, and self-esteem. You may desperately want to remain with a problem and find that you cannot. Nothing here diminishes that. Nor does it diminish the value of a serene mind. The ability to hold a difficult problem steadily can be extraordinary. Some lives are enlarged by deep concentration. Some problems require years of patient work.

The question is not whether sustained attention is good. It is whether every departure from an externally preferred state should be interpreted first as failure.

Suppose the externally imposed sequence says:

$$P_A \to TS_A.$$

You repeatedly move elsewhere. The observable behavior is real. But several explanations remain possible.

Perhaps you cannot control attention in a way you yourself find harmful. Perhaps the assigned problem is insufficiently salient or conflicts with an active criticism. Perhaps another problem has become more urgent, or the present state has already been exhausted for you. Perhaps the difficulty is physiological. Perhaps several mechanisms interact.

The behavioral surface does not decide among them. That is the epistemic point.

Attention-deficit/hyperactivity disorder makes the distinction especially important. The category refers to real patterns of difficulty and, for many people, real suffering.

Interventions can help. Medication can help. Behavioral and environmental changes can help. Successful treatment matters.

Compare it:

$$\text{treatment efficacy} \neq \text{complete explanation.}$$

If a medication helps you remain with work you care about, that is important knowledge about the intervention. It does not, by itself, prove why your prior experience had the structure it did.

Observations of arrows offer no clinical explanation of ADHD. It suggests an epistemic conjecture that may reveal how much remains compressed inside the observable category.

Imagine being required to remain in an epistemic state after the live arrows of your mind have moved elsewhere. As the imposed state becomes less inhabitable, concentration on it deteriorates. Other transitions keep surfacing.

From outside:

> inattentive.

You are required to suppress those transitions. You become restless. The movement surfaces behaviorally.

From outside:

> hyperactive.

Again, this is only a conjecture. It cannot explain the clinical condition. It may be false. Its purpose is narrower: to show that the same observable pattern can be compatible with an epistemically richer interpretation than:

> the person fails to sustain proper attention.

That possibility is morally relevant because it changes what we believe may be happening inside the person we are judging. The same caution applies to treatment itself. Psychological categories can become most convincing when treatment works. The reasoning can become:

> this person has difficulty concentrating; medication improves concentration; therefore the diagnosis has captured the mechanism; therefore the prior mode of attention was defective in the way the diagnosis implies.

The last two moves do not follow automatically. A painkiller can relieve pain without revealing the cause of the injury. A breathing technique can reduce panic without

proving that every panic episode has the same explanatory history. A stimulant can alter cognition without explaining the full epistemic structure of the person before treatment.

Efficacy is not ontology. Useful technology does not automatically settle the explanatory question. You may be genuinely better off. That can be true without converting the intervention into a theory of personhood.

This is not skepticism about interventions. It is a reminder that successful intervention not be asked to perform explanatory work it has not earned.

6. Fast and Slow Thinking, Reversed

A familiar distinction also looks different from the perspective of the arrows. We often contrast fast and slow thinking. Fast cognition is associated with immediacy, intuition, heuristics, and automaticity. Slow cognition is associated with deliberation, reflection, and care.

Whatever empirical distinctions lie beneath those descriptions, the ordinary interpretation usually assumes a relatively stable task or problem frame. You are thinking fast or slowly **about this**. Observations of arrows adds another dimension.

Suppose you remain inside a problem for a long time:

$$P_A \rightarrow \text{processing} \rightarrow \text{processing} \rightarrow \text{processing}.$$

To an observer sharing the frame, this looks like slow, careful thought.

Now consider:

$$P_A \rightarrow TS_A \rightarrow EE_B \rightarrow P_C \rightarrow TS_D.$$

The transitions may occur rapidly. Your epistemic mind is moving. But the observer has to rebuild the frame at every step. The observer experiences:

> slow down.

So the conventional intuition can invert.

> **For the transitory thinker, rapid state transition can feel fast while appearing slow to the state-anchored observer.**

And:

> **For the state-oriented thinker, sustained processing can feel appropriately slow while appearing painfully slow to someone whose arrows have already moved.**

The difference reveals an anchoring bias. We measure cognitive speed against the state we understand. When another person leaves that state, some of what we call their slowness may be our own reconstruction cost.

This does not eliminate useful distinctions between impulsive and reflective cognition. It makes the moral ranking less secure. Legibility is not intelligence. Nor should an argument about arrows romanticize movement.

Some people love deep continuity. They can remain with one problem for hours, years, even a lifetime. The arrow may point inward rather than away: a tentative solution becomes a world of criticism and refinement; routine produces skill; discipline preserves a problem long enough for something difficult to become possible.

The point is not:

> movement good; stasis bad.

It is:

> **Your epistemic direction should not be inferred from an externally preferred pattern of movement or stillness.**

Freedom of arrows includes the freedom to remain.

7. The Moral Weight of Invisible Arrows

A familiar compression now deserves the same treatment as the epistemic cycle.

All life is problem-solving.

Life persists through solutions to problems. Organisms embody knowledge. People create explanations.

But the grammar favors the nouns. Problem. Solving. We picture a problem state and a solution state. The hyphen disappears.

Perhaps the hyphen matters as much as either side. Problem-solving includes recognizing, trying, criticizing, abandoning, returning, reframing, becoming interested or bored, and discovering that the problem itself changed. The hyphen is movement.

If epistemic states are places you temporarily inhabit, the arrows are part of the journey of your epistemic life. A life cannot be reduced to its destinations. Neither can epistemology.

The journey is not an error between destinations. The transition is not empty space between achievements. It may contain much of the inexplicit transformation from which the achievement becomes possible.

Abstractly:

$$P_1$$

is not the same epistemic object as:

$$P_1 \rightarrow TS.$$

The arrow remains mysterious and compressed, but it exists.

The hidden depth of people should change how we interpret them. The visible person is a shallow interface. Behind it lies a vast and largely inaccessible epistemic mind. Arrows add a dynamic consequence. The movements inside that mind deserve moral consideration.

Suppose an arrow carries an emerging criticism, a new problem becoming salient, a connection not yet explicit, the impulse to leave an exhausted problem, or the need to remain with a difficult one. Neither you nor the observer may yet be able to explain it.

If the movement is interrupted, perhaps nothing important was lost. Perhaps the interruption itself creates the better problem. But perhaps a real epistemic transition was interrupted.

We do not know. That ignorance should matter.

The conclusion is not:

> never interrupt.

Life is interruption. Reality interrupts us. Criticism interrupts explanations. Other people expose problems we did not choose. Some interruptions are the arrow. The stronger concern arises where an authority systematically decides that only externally prescribed transitions are legitimate.

The burden then shifts from:

> why should we allow this movement?

Toward:

> what explanation makes suppressing it necessary?

This is not a presumption that every movement is valuable. A transition can lead toward error. A mind can cycle without development. A harmful explanation can redirect an entire life. A self-sealing worldview can turn every criticism into another reason to remain inside it.

Fun does not certify truth. Movement does not certify progress. The epistemology remains fallibilist. Arrows deserve criticism precisely because they are mysterious and ubiquitous.

8. Other Minds Are Expensive to Follow

Why do institutions prefer stable states?

One reason is simple. Transitions are costly for other people. A shared problem frame compresses communication. Everyone knows the terms. The project has a scope. The procedure has steps. As long as everyone stays in the state, coordination is cheap.

Now let one person move. They question the problem. Change the criterion. Introduce another domain. Reject the premise. The others must reconstruct. The transition may be valuable. It may also consume enormous attention.

Historically, one solution has been to require the transitory person to become more state-like.

Stay on topic. Be concise. Follow the process. Finish the task. Do not reopen the question. Don't overthink it.

That can be a reasonable local coordination rule. It becomes dangerous when the coordination cost is converted into an ontology:

> this person's way of thinking is defective because it is expensive for us to follow.

The burden of legibility has been placed entirely on one side.

Technology may allow another arrangement. The human-AI dyad is a way to extend your epistemic reach. AI can help you externalize, translate, preserve, compare, reformat, decompress, and explicate.

Arrows suggest another use:

> **AI can compact the trail of epistemic arrows.**

Imagine working through a problem over days. You jump between notes. Return to old arguments. Introduce examples. Reject them. Recover something from another project. Change terminology. Reformulate the problem. To another human following synchronously, the trajectory may be exhausting.

AI can preserve much more of the evolving context. It can recover earlier states, make transitions explicit, compact repetition, and reorganize a highly dynamic process into a stable artifact.

You do not need to become more state-like. The interface becomes better at preserving the movement.

This does not make the AI the originator of your epistemic direction. It performs an objective-knowledge function. You move. The technology helps preserve the trail.

Abstract it.

This can transform a difficult-to-follow epistemic journey into something another mind can reconstruct:

$$M_A \rightarrow \text{epistemic movement} \rightarrow O \rightarrow M'_B.$$

The objective artifact O may be an essay, a diagram, a piece of software, or a theory.

The recipient does not have to reproduce the creator's entire journey. That has always been true of objective knowledge. AI can make the transformation from private movement to stable surface cheaper and more continuous.

That may have a moral consequence. People whose cognitive transitions were historically expensive for others to follow may gain new ways to make those transitions legible without suppressing them. The playing field changes not by normalizing the person but by improving the bridge.

9. Institutions Route the Arrows of Persons

Groups do not possess a collective subjective arrow. Society does not think. The corporation does not become curious. The bureaucracy does not experience boredom.

People do.

Institutions nevertheless create structures that route our arrows. They specify which problems count, who may define them, which tentative solutions are admissible, what criticism is permitted, when the discussion is over, and which successor problems receive resources.

This is unavoidable to some extent. Coordination requires interfaces. A hospital cannot allow every worker to reinterpret a safety protocol during an emergency. A flight crew cannot spontaneously reinvent landing procedure. A factory cannot permit every operator to change product specifications. Constraints can enlarge freedom by making complex cooperation possible.

The epistemic criterion remains:

> which constraints make more independent explanatory lives possible?

Arrows suggest a dynamic version:

> **Which constraints coordinate necessary states while leaving arrows open?**

Top-down systems are attractive because they simplify coordination. The authority defines the problem, selects the solution, assigns roles, specifies sequence, and measures compliance. Where the required transformation is already known, this can work extremely well. Known transformations can be standardized, and objective knowledge can be embodied in procedure.

But knowledge creation introduces another problem. The authority does not merely need the correct state. It must know where the system should move next. That requires knowledge the authority may not possess.

Worse, the relevant knowledge may not yet exist. A worker may notice a problem no manager has formulated. A scientist may challenge the premise of the research program. A dissenter may expose a problem in the official explanation.

Command-and-control can therefore fail for a reason deeper than incomplete information. It can suppress the process by which missing knowledge would have been created. It imposes not only static structure on a dynamic epistemic landscape. It can impose static arrows on the persons capable of changing that landscape.

Organizations often recognize the resulting disorientation. Employees seem disengaged. Teams lack alignment. People resist change.

The proposed solution is frequently more communication. Leadership must explain the reality, share the facts, clarify priorities, align everyone around the problem. Sometimes this is exactly what is needed. Miscommunication is real. Missing objective knowledge can create unnecessary conflict.

But the solution can reproduce the original mistake.

It assumes:

$$\text{shared information} \rightarrow \text{shared problem} \rightarrow \text{shared direction}.$$

The same objective knowledge enters different M_i. People reconstruct it through different histories, problems, skills, obligations, fears, interests, and inexplicit criteria.

Corporate facts do not comprise a corporate mind. More information can improve coordination without producing the arrows management wants. When leadership interprets continued divergence as failure to understand reality, the system begins to confuse its own problem formulation with reality itself.

The same issue appears in economic and political institutions. Free markets remain an interesting epistemic conjecture not because markets are infallible, prices explain everything, or economic value equals epistemic value. Markets can concentrate power. Exit can be expensive. Competition can reward triviality. Firms operating inside markets can be profoundly hierarchical.

The narrower observation is that decentralized economic systems can permit many problem situations and many arrows to coexist. One person sees a problem, another sees a different one, a third builds a substitute, and another changes the problem entirely. No central authority needs to know in advance which transition will matter.

The epistemic content of the arrows strengthens an old criticism of command economies. The difficulty is not merely that the planner lacks distributed information. The planner lacks **uncreated knowledge**. If the planning system tightly controls which problems may be pursued, which experiments may occur, and which criticisms may alter direction, it can impair the generation of the very knowledge it lacks.

This does not prove liberal economics. It identifies another point of consilience between decentralized institutions and an epistemology of fallible, problem-generating persons.

The same idea applies more broadly to liberal institutions. Freedom of speech matters partly because criticism must be able to enter. Freedom of association matters partly because people need to form new problem-solving arrangements. Exit matters because you may need to leave one imposed problem situation for another. Freedom of conscience matters because inherited explanations must remain criticizable inside you. Political liberty matters because authorities are fallible.

Liberty is therefore not merely permission to occupy different states. It is also permission to move.

To change your mind. To abandon a theory. To form a new problem. To pursue an interest others find strange. To criticize the official sequence. To become someone whose future problem situation was not predictable from the present one.

That does not eliminate coordination. It changes the problem from:

> how do we make everyone follow the right arrows?

To:

> how do we build interfaces through which many different arrows can coexist without destroying the shared conditions that make movement possible?

10. The Open Future Arrives One Arrow at a Time

The failure of forced arrows is not only economic. It can be psychological, moral, and creative.

If you are repeatedly required to follow prescribed transitions, you may become less able to recognize your own. The signals become inconvenient. Boredom must be ignored. Curiosity must wait. Criticism must be softened. The alternative problem must be suppressed.

Eventually you may become good at staying in the permitted state. That can look like maturity, professionalism, discipline, or reliability. Sometimes it is. Sometimes it is the successful internalization of epistemic control.

The distinction cannot be read from compliance alone. An institution may obtain predictable behavior while reducing the number of independent explanatory trajectories available inside it. That tradeoff may be necessary. It may also be far more expensive than the institution can see.

Lost knowledge does not produce a report. The problem that was never formed leaves no metric. The conjecture that never became explicit cannot appear in the retrospective.

Suppressed arrows are difficult to count because their products do not exist. That is precisely why they are hard to value.

Freedom preserves an open future. Arrows give that abstraction a local form. The open future is not encountered all at once. It arrives through proximate transitions. This problem becomes that conjecture. This criticism becomes that new problem. This interest becomes that skill. This mistake becomes that explanation.

Your future remains open because the next arrow is not fully specified by the present state. Causal history, biology, objective knowledge, institutions, and physical possibility all matter. But the explanatory future is not already contained as a semantic script inside any of them.

> The next problem may not yet exist. The next criterion may not yet exist.
> The next explanation may not yet exist.

The next person you become may depend upon arrows nobody could presently name.

That is one reason control should make us epistemically cautious. To prescribe every arrow is to presume knowledge of a future that has not been created.

The same point appears in the ordinary experience of being stuck. You cannot think. Nothing is moving. So you stop trying to force a destination. You walk. The mind drifts. A memory appears; something unrelated becomes irritating; a connection forms. Perhaps nothing useful happens. Perhaps the entire afternoon is wasted. Or perhaps the successor problem appears.

The important point is not that wandering is secretly efficient. That would convert freedom into another productivity technique.

The point is that sometimes you know only this:

> **The current state is not where I should remain.**

That is already an epistemic orientation. The arrow can become consilient before the destination does. Movement itself can be the next thing that makes sense.

11. Coordinate the Interface, Not the Arrows

This clarifies why the institutional problem is so difficult. A command system is not merely trying to coordinate many people. It is trying to coordinate many moving epistemic worlds.

Abstract it:

$$M_1(t), M_2(t), \ldots, M_n(t).$$

Each epistemic mind contains different problems, histories, inexplicit structures, transitions, rhythms, and ways of becoming stuck or interested and of externalizing what has changed.

The system then imposes:

$$P_A \to TS_A \to EE_A \to P_B.$$

The mismatch is structural. A static representation of the organization is applied to dynamic explanatory agents. The more complex the problem environment becomes, the more expensive that mismatch can become.

The irony is that the system may respond by increasing control: more process, more reporting, more standardized problem definitions, more prescribed arrows. The attempted solution can intensify the original epistemic failure. People become easier to administer and less able to generate the knowledge the system cannot anticipate.

Perhaps coordination does not require a mind above minds. Perhaps it requires better bridges among them. Arrows strengthen the point. The bridge must accommodate movement.

It is not enough to connect static representations of people. People will change. Their problems will change. Their criteria will change. Their explanations will change. Their next useful direction may not yet be knowable.

Objective knowledge can help. Contracts can define thin interfaces. Standards can stabilize what must remain shared. Markets can coordinate differently motivated agents. Liberal institutions can preserve criticism and exit. AI can reduce the cost of following complicated epistemic trajectories.

None solves the problem. But all suggest a direction:

Coordinate the interface without trying to own the arrows.

That may be one of the deepest design principles available to a civilization of solitary, changing explainers.

12. The Arrows Between People

There is also a gentler version of the same problem. One person speaks. Another listens. The first person's explanation moves somewhere unexpected. The second becomes lost.

Ordinarily, this is frustrating. But if you are to take another person's explanation seriously, sometimes you must reconstruct not only the state but the transition.

Why did that objection follow? Why did that memory become relevant? Why did they abandon the previous line?

You may never know fully. The road through another M_i remains partial. But asking about the arrow changes the interaction.

Instead of:

> you're off topic,

you can sometimes ask:

> what moved for you there?

Not because every transition deserves indulgence. Because it may contain the problem you are missing.

This is not only intellectual. To know another person over years is partly to learn their arrows: the subjects that light them up, the problems they cannot release, the criticisms that wound, the bait that rages, the silences that mean a transition is occurring, the need to remain, the need to leave.

You still do not occupy their mind. The solitude remains. But familiarity can reduce the reconstruction cost. A loved one's strange transition becomes less strange because you know more of the landscape.

Perhaps this is one meaning of being known. Not that another person possesses your internal state. They begin to recognize your arrows. A form sees the state. A relationship can sometimes intuit the arrow.

13. What Arrows Do Not Explain

The abstraction of arrows should not be mistaken for biology. Nothing here establishes that the brain contains discrete epistemic components connected by literal transition mechanisms corresponding to the notation.

The states are abstractions. The arrows are abstractions. Even the distinction between state and transition may carve the underlying physical process in a way that is useful for epistemology without mapping neatly onto biological machinery.

That is acceptable. An explanation need not instantiate the vocabulary of a different level. But it must not pretend the physical mechanism has already been explained.

So the position remains deliberately modest:

> people appear to move among distinguishable epistemic states; those movements carry epistemic significance; the causal machinery remains inexplicit and unexplained.

The abstraction earns its place if making the movement visible creates better problems.

Observations of arrows has not explained the causal machinery that moves people among epistemic states. It has not explained attention, ADHD, creativity, or fun. It has

not shown that decentralized systems always outperform centralized ones or that every spontaneous transition deserves protection.

You remain vast and solitary. Other people reconstruct you incompletely. Institutions see interfaces. Even you may notice the direction only after it has begun to change. But the transition need not be treated as empty space between the states we name.

There may be knowledge in the arrow. And sometimes, the only thing you know is that it is time to move.

So how do we know where to go next?

If future knowledge depends on creative transitions that have not yet occurred, what can we possibly know about the person you will become?

XVIII. Unpredictable You



Creativity appears throughout the epistemic cycle: in problems, tentative solutions, criticism, successor problems, and the transitions among them.

Fallibility travels with it. A creative problem can be a bad problem. A creative criticism can be mistaken. A transformed evaluative criterion can make an explanation worse. A novel transition can move you into a more epistemically closed world.

Creativity is not a guarantee of progress. It is the capacity to bring new explanatory structure into the world.

That capacity suggests a problem: If genuinely new explanatory knowledge can appear, then the future epistemic landscape does not yet exist. We can know an enormous amount about the present. But the knowledge that will change the world is not yet available.

The problem is therefore deeper than uncertainty. The future is not merely hidden; it is open. And if people create the knowledge from which later history will be made, then we are not merely living inside an unpredictable world. We are among the causes of its inherent unpredictability.

1. Creativity Is Distributed Through the Cycle

The schematic can tempt us into assigning creativity to one place:

$$P_1 \rightarrow \mathbf{TS} \rightarrow EE \rightarrow P_2.$$

The problem is given, you invent a solution, the solution is criticized, and the cycle continues. That picture is too simple.

You can live for years inside an inherited problem and then realize that the problem was badly formed. The question changes. The evaluative criteria change. What once counted as progress becomes irrelevant. You have not merely produced another answer; you have created a new region in which answers can exist.

Criticism can do the same. It need not merely eliminate a tentative solution. It can reveal that the original opposition was false or expose an assumption that neither side had noticed. It can also connect two previously distant explanations and create a contradiction that did not previously exist for either participant. The successor problem is then not simply the old problem with one bad answer removed. It is new knowledge.

The arrows matter too. You can move from one problem to another without a complete explicit account of why. A memory becomes relevant, a distinction becomes salient, an old irritation becomes a problem, or something that had seemed settled becomes unstable. We do not yet possess a complete causal explanation of these transitions. But we can already reject one simplifying picture: creativity is not a module that activates when you need an idea. Nor is it a special mode reserved for artists, scientists, inventors, or children. It is implicated in the entire growth of knowledge.

That also means creativity does not always lead to progress. You can invent a better explanation or a better rationalization. You can create a criticism that exposes error or one whose function is to prevent another criticism from being heard. You can create a moral category that enlarges sympathy or one that makes exclusion easier. You can invent a tool for freedom or a tool for surveillance. The epistemic capacity is the same. The direction differs.

2. The Future Knowledge Problem

Suppose someone asks:

> What will people know fifty years from now?

The question sounds ordinary. We may list likely technologies, extrapolate scientific trends, forecast political arrangements, or describe future cities, medicine, education, artificial intelligence, energy systems, relationships, and institutions. We may give probabilities or construct scenarios.

But the deepest problem appears before any of those exercises begin. If the future knowledge is genuinely new, then we do not yet possess the explanation that would make it available for prediction. To specify the explanatory content of the future discovery now would be, in the relevant sense, to have already created it.

This is not merely a shortage of data, insufficient computing power, or the complexity of the future. The future explanation does not yet exist as objective knowledge. That is why the growth of knowledge is intrinsically unpredictable.

A new explanation has a causal history. You can struggle with a problem for months while your skills change, your attention shifts, a criticism weakens an old explanation, and a new abstraction slowly becomes available. Several apparently unrelated regions of your epistemic mind can change incrementally. Then you say:

> I just realized something.

The explicit appearance can be discontinuous even when the causal preparation is not.

The same structure can appear outside your solitary mind. A scientific revolution may later be narrated as a leap. A technological platform may seem to arrive suddenly. A moral transformation may be attached to one speech, one book, one court decision, one protest, or one invention.

Yet the visible event can emerge from countless smaller changes distributed among many loosely coordinated persons. Someone invents a tool. Someone else develops a notation. Someone elsewhere creates a criticism. An institution changes, a technical constraint disappears, a concept becomes culturally legible, or a problem that once seemed local becomes general. Eventually one visible artifact reorganizes the landscape.

The historical surface looks discontinuous because the causal history is distributed. Openness therefore does not require magic. It requires only that genuinely new explanatory structures can be created through causally continuous processes.

3. How Prediction Closes the World

The impossibility of predicting future knowledge is easy to state and difficult to inhabit because common sense pulls in the other direction. We look backward, see patterns, extend them, and call this experience. History seems to teach us how history behaves.

The familiar phrase is:

> History does not repeat, but it rhymes.

The phrase denies crude repetition, but the deeper epistemic structure remains:

$$\text{past pattern} \rightarrow \text{future pattern}.$$

The rhyme is still supposed to be available in the past. The future remains a transformed continuation of an already represented possibility space.

Modern predictive reasoning can look as though it escapes that structure. Instead of saying, "This happened before, therefore it will happen again," we now say something more precise: "This happened before, so the posterior probability should increase." The formal tools are more powerful too. The epistemic move can nevertheless remain the same: represent a possibility space, assign probabilities within it, and use new observations to update those probabilities.

This can be extremely useful inside a closed enough world. The problem begins when the representation is promoted into an account of knowledge creation itself. A probability distribution operates over represented possibilities, while knowledge creation can alter what the possibilities are. You can create a new option, reject the original categories, or redefine the problem. You can invent a criterion that makes the previous ranking meaningless or produce an explanation showing that the original possibility space was badly formed.

The progression becomes more sophisticated: from "history repeats" to "history rhymes," then to probabilistic resemblance and updating credence as confirming observations arrive. The problem remains.

The same inversion appears when a useful model becomes a theory of the mind using it. The Bayesian model becomes the Bayesian brain. A person who can employ a probabilistic representation is redescribed as though their epistemic life consists of assigning and updating weights over an already specified world. The categorical error: a psychometric compression becomes personality, a diagnostic compression becomes the person, a statistical regularity becomes a causal ontology. Now a probabilistic model of represented uncertainty becomes a model of knowledge creation itself.

The tool is real, and so is its usefulness. The promotion is the error.

Once a projected future exists, observations can begin to accumulate around it. A predicted trend continues, so confidence increases. An event resembles the scenario, so the model appears corroborated. A contrary observation arrives and is classified as noise, temporary deviation, or an exception anticipated by another branch of the model. The representational structure grows more elaborate, and its apparent credibility increases because it can absorb more of what happens.

Abstract it.

The sequence can become:

extrapolation → future model → credence → confirming observations → greater credence.

This is deeper than confirmation bias as a psychological tendency. The future model has been granted the epistemic status of a space within which later observations are interpreted. The closed world becomes the frame. You then move through imagined

histories, updating priors and counting confirmations inside a representation whose most important omission is the knowledge that does not yet exist.

Such a model may be mathematically sophisticated. It may generate precise outputs, survive many observations, and support useful decisions. None of that changes the status of the represented possibility space itself: it remains conjectural. A new technology can invalidate the category by which a trend was represented. A moral criticism can transform what later people regard as an acceptable outcome. A new institution can change the relevant action space. A scientific explanation can show that the variables in the original model were badly chosen.

The failure, then, need not be that the forecast was slightly wrong. The represented future can cease to be the relevant future.

4. You Are Not a Projection of Your Present Compression

The argument becomes more intimate when we return to you and your solitary mind. That mind includes explicit explanations, but also inexplicit knowledge, memory, skill, salience, affect, expectation, perceptual organization, concepts, language, social knowledge, unfinished problems, and changing evaluative criteria. No outside observer possesses direct access to this vast universe, and you do not possess explicit access to it either.

Yet we regularly compress the epistemic mind, M_i. We call someone conscientious, anxious, agreeable, gifted, difficult, traditional, impulsive, depressed, creative, uncreative, a leader, a follower, a risk, or a type. Some of these compressions can be useful. The epistemic error appears when compression becomes projection:

$$M_{i,t_0} \to \hat{M}_{i,t_0} \to \hat{M}_{i,t_1} \to \hat{M}_{i,t_2}.$$

You are first made simple in state and then simple in evolution. Your future is predicted from the compression.

The problem is not only that categories fail to capture everything about you. **The dimensions discarded by the category may include the very knowledge creation through which you become different.** You become predictable in the model because the model first makes you smaller—not smaller in reality, but smaller in the representation.

The moral consequence becomes clearer when the compression concerns another person. Imagine that I construct an explanation of them: selfish, unreliable, ideological, anxious, narcissistic, timid, ambitious. Perhaps the description captures something observable. But once the category becomes stable, it begins to travel forward. A generous act becomes strategic. A disagreement becomes diagnostic. A change of mind

becomes inconsistency. A surprising action becomes evidence for a deeper version of the same trait.

The model begins to absorb its own criticism. The other person becomes increasingly predictable because the explanation has become increasingly difficult to escape. The problem is not merely that I may be wrong. I have reduced the other person's epistemic openness before encountering it.

A broader approach begins by keeping the explanation exposed to failure. Suppose the person acts in a way I do not understand. Instead of asking:

> What kind of person behaves like this?

I can ask:

> What problem would make this action intelligible?

That question produces several tentative reconstructions:

$$A \to \{P_1, P_2, P_3, \dots, P_n\}.$$

Perhaps the person is solving a status problem, a safety problem, or a loyalty problem. Perhaps they are responding to a moral problem I do not yet recognize. Perhaps my description of the action is itself wrong. One of these models need not secretly be the true one. Their value here is that the wider representational surface creates more opportunities for my explanation to fail.

This is **the rationality principle** used as a surrogate rather than a prophecy. The model says:

> If this were the person's problem situation, the action would make sense.

It does not say:

> Therefore this is what the person will do next.

Its epistemic value comes from becoming criticizable. The observed person remains able to destroy the reconstruction. To explore another person's mind in this way is therefore also a moral practice. It refuses premature closure and treats the other person as a universe that cannot be exhausted by the compression through which they first became legible to us.

5. How Openness Gets Recruited Against Itself

Children make epistemic openness difficult to ignore. A child can turn a chair into a spacecraft. A question about an insect can become a question about death. A game can change its criteria halfway through. A child can ask a question that exposes an assumption adults stopped noticing years ago.

The claim is not that childhood is a mystical state of pure creativity. Children are fallible. Their problems can be confused, their explanations wrong, their criticisms arbitrary, and their inventions unable to survive contact with reality. What childhood makes conspicuous is open-ended epistemic movement.

Coordination changes the environment in which that movement occurs. Schools need schedules, curricula, assessments, standards, classrooms, roles, and expected outputs. Organizations need employees to perform identifiable tasks. Professions need credentials. Bureaucracies need forms. Coordination therefore requires compression.

A child who freely creates problems enters environments in which the problem is often supplied in advance. The solution space narrows, the evaluation criteria become externally legible, and the transition schedule is imposed.

Abstract it.

The sequence can become:

assigned problem → permitted solution → authorized criticism → next assigned problem.

This does not eliminate creativity. If it did, communication itself would become impossible. You still have to interpret instructions, reconstruct meanings, resolve ambiguities, form expectations, and integrate new information into an epistemic mind no institution possesses. The creative process continues, but its direction can be constrained and recruited.

Creativity can even be used to impede further creativity. A manager can invent a more sophisticated performance system that makes dissent harder. A bureaucracy can create increasingly elaborate categories through which people must become legible. A political movement can create a vocabulary that turns criticism into evidence of disloyalty. An institution can create procedures whose complexity makes responsibility difficult to locate. Objective knowledge can preserve and scale these creative attempts to suppress further creativity.

The same structure can operate inwardly. You encounter an inherited explanation of what a successful person should be. You do not merely submit mechanically; you can creatively reconstruct yourself within it. You explain why your interests are impractical and reinterpret dissatisfaction as immaturity. You explain away a problem that threatens an identity and invent reasons that the life expected of you is the life you really wanted. The resulting self-explanation may become extraordinarily elaborate.

Creativity has not disappeared. It has been recruited into epistemic closure.

That is why saying "I am not creative" can distort what is happening. If personhood involves creating problems, explanations, criticisms, interpretations, and transitions among epistemic states, then artistic creativity is one conspicuous expression of a more general epistemic capacity. The issue is not that everyone is secretly an artist. It is that your epistemic life already depends upon forms of creation that a narrow cultural category of creativity may fail to name.

6. The Geometry of Openness

The distinction between compression and openness now recurs at several scales. The error has the same shape each time: compress an open present, project the compression, and mistake continuity in the model for knowledge of what will become.

Abstract it.

At the level of civilization:

$$K_{t_0} \to \text{compress} \to \text{project deeply}$$

vs:

$$K_{t_0} \to \text{expand broadly} \to \text{seek contradiction.}$$

At the level of a field:

$$E_{t_0} \to \text{narrow model} \to \text{forecast}$$

vs:

$$E_{t_0} \to \text{connect explanations} \to \text{consilience} \to \text{new problem.}$$

At the level of a person:

$$M_{i,t_0} \to \text{trait / category / profile} \to \text{behavioral prediction}$$

vs:

$$M_{i,t_0} \to \text{multiple tentative reconstructions} \to \text{criticism} \to \text{deeper encounter.}$$

At the level of the self:

$$M_{\text{self}} \to \text{identity compression} \to \text{expected future self}$$

vs:

$$M_{\text{self}} \rightarrow \text{new problems} \rightarrow \text{new criticism} \rightarrow \text{open future self.}$$

That recurring geometry sharpens the question we can ask of objective knowledge. At t_0, our representational surface is extraordinarily broad. We possess objective knowledge about physical systems, biological processes, technologies, institutions, laws, markets, languages, software, mathematics, scientific explanations, engineering systems, historical records, norms, organizational processes, and innumerable local causal structures.

The breadth is vast, but the depth of legitimate causal projection into an epistemically open future is comparatively shallow. We can represent many things near the present. We cannot represent the future knowledge that will later reorganize them. The epistemic landscape may therefore look less like a narrow road extending far into the future. It may look more like an enormous surface lying close to the present: very broad, relatively shallow.

Much of our representational activity points in the opposite direction. We build narrow explanatory surfaces and project them many steps forward. Economic forecasts, political forecasts, technology forecasts, demographic forecasts, market forecasts, risk scenarios, strategic roadmaps, and histories of the future all take this form. The representation has depth; the causal basis may not. The farther the projection travels through domains saturated by human knowledge creation, the more opportunities there are for explanatory novelty to alter the path. The future representation grows thick while the epistemic foundation grows thin.

Consilience points across the broad surface instead. If there is one reality, explanations developed in different domains cannot remain permanently incompatible. A biological explanation can collide with a psychological explanation. A computational explanation can collide with a theory of mind. A moral explanation can collide with an institutional practice. A statistical model can collide with an explanation of the person. The collision creates a problem.

This is one reason creating a genuinely good explanation is difficult. An explanation may look excellent inside the narrow surface where it was invented; its weakness becomes visible only when its consequences reach another domain. The breadth of objective knowledge therefore creates an enormous error-elimination surface. More knowledge does not merely mean more facts. It means more opportunities for explanations to collide.

The direction of epistemic progress is therefore not necessarily:

$$\text{deeper projection.}$$

It may be:

$$\text{greater breadth} \rightarrow \text{greater consilience} \rightarrow \text{more contradiction} \rightarrow \text{better problems.}$$

This is not a retreat from explanation. It is a different use of explanatory reach. Present categories receive less authority over the distant future, while present explanations face more pressure to survive contact with everything else we already know.

7. AI Depth or Breadth

Modern computation gives us extraordinary representational scale. A machine can retrieve, transform, compare, and preserve more objective knowledge than one biological mind can hold explicitly, and artificial intelligence extends that scale dramatically.

One obvious use of that capacity is to project represented structure forward. We ask machines to forecast, rank, score, predict, estimate, assign probability, generate scenarios, project trends, and model future behavior. The machine receives a possibility space, extends it, assigns weights, elaborates consequences, and generates a future that appears increasingly detailed. This can make epistemic thinness look like knowledge.

A model begins with a few assumptions. The machine produces ten pages of consequences. The output is coherent, specific, internally connected, perhaps quantitative. The original assumptions disappear beneath the representational richness. A thin future looks thick.

Generative AI is unusually well matched to this operation because continuation is one of its conspicuous abilities. Give it a sequence and it can continue the sequence. Give it a trend and it can extend the trend. Give it a set of political assumptions and it can generate future political scenarios. Give it a technological roadmap and it can populate the decades around it. Give it a profile of a person and it can produce imagined future behavior.

The output can be compelling because it resembles the structure of historical explanation. Events lead to events. Causes accumulate. Institutions respond. People adapt. Technology advances. The future becomes narratively legible. But narrative legibility is not future knowledge.

The machine has extended inherited structure. It has not thereby created the explanatory knowledge that future persons will later create. This is the external analogue of predicting the compressed person: a narrow representation is projected forward, the generative system fills the empty surface, and openness disappears inside prose.

Probabilistic machinery can make the projection look firmer still. Scenario A receives 50%, Scenario B 30%, Scenario C 20%. The imagined future now has numerical credence as well as narrative detail. New events can be interpreted as movement toward one scenario or another, and the future is progressively redescribed as though it had always been contained in the original model. The history of the future is closed.

But computation need not be used primarily for depth. A model can make an explanation precise, expose a dependency, show what would follow if a conjecture were

correct, and make criticism easier. The rationality principle is useful for exactly this reason. A model of another person can say:

> If this were their problem situation, this action would make sense.

Reality can contradict that reconstruction. The same logic can operate at larger scales. A model of an institution can say:

> If these constraints are operative, this outcome should follow.

A model of a technology can say:

> If this causal architecture is correct, this intervention should produce that effect.

The model is useful because it creates a surface for error elimination. The mistake is to promote that surface into prophecy.

This changes the computational problem. Instead of asking which represented future is most probable, the machine can help expose which presently available explanations cannot all be true together. Instead of predicting what another person will do, it can generate several problem situations that would make an action intelligible. It can then identify what would disconfirm each reconstruction. Instead of asking which technology will dominate in 2040, it can expose the present constraints, explanations, and unresolved contradictions that make current trajectories unstable. Instead of narrating the future of education, it can place explanations of learning, authority, childhood, measurement, and knowledge creation beside one another and reveal where they conflict.

This does not make the future more predictable. It makes the present more criticizable.

Artificial intelligence may be particularly useful here because biological cognition is narrow at the explicit surface. You cannot simultaneously hold every relevant theory, paper, institutional rule, historical example, technical constraint, philosophical implication, and conflicting explanation. Objective knowledge already extends you beyond those limits: writing extends memory, mathematics extends representation, libraries extend retrieval, scientific institutions extend specialization, and computers extend calculation.

AI can extend comparison. It can place explanatory structures beside one another that no individual would otherwise hold together long enough for criticism to operate. It can help reveal hidden contradiction and generate alternative rational reconstructions. It can expose where a model depends on a category inherited from somewhere else, or where an explanation in one field quietly requires something another field says is impossible.

That is consilience as an engineering problem. You remain the epistemic locus; the machine enlarges the surface on which problems can appear.

The unpredictability of knowledge therefore suggests a research program. How can large bodies of objective knowledge be represented in ways that preserve explanatory structure rather than merely textual similarity? How can machines expose contradictions across fields? How can they generate criticizable rational reconstructions without converting those reconstructions into predictions? How can they preserve multiple candidate explanations without collapsing them into a probability distribution? How can they help you discover problems rather than merely answer inherited ones? How can they make conflicting evaluative criteria visible? How can they show where an apparently stable future depends on the assumption that nobody creates a new explanation?

These are technological questions, but they arise from the epistemology.

8. Planning Without Possessing the Future

The inability to know the future is not a new human problem. Traditions have located the future in fate, providence, prophecy, divine will, historical necessity, technological progress, economic law, and statistical expectation. The explanations differ enormously, but many can be reconstructed as responses to the same problem:

> How should you act when what happens next cannot be known?

One recurring response is to create an authority with privileged access to the future. The role can be occupied by the oracle, the priest, the ruler, the historian of destiny, the expert forecaster, the probabilistic model, or the algorithm. The forms differ, but the attraction is similar. If the future can be known by someone or something, uncertainty becomes governable.

The epistemic problem returns wherever future history depends upon explanatory knowledge not yet created. No authority possesses that knowledge in advance. An authority may still know more about some present causal structures, construct better models, or identify real constraints. None of these achievements converts open knowledge creation into a closed future.

This does not make expertise useless or planning incoherent. You still act. Institutions coordinate. Engineers design. Companies allocate resources. Governments build infrastructure. Parents make decisions. The openness of knowledge changes the epistemic status of those activities rather than eliminating them.

A plan is a tentative solution. A roadmap is a conjecture. A scenario is a model. A forecast is a representation. Each can be useful and each can be criticized. None contains the future.

Planning under openness therefore has a different architecture. Action proceeds from the best present explanation while alternatives remain available, assumptions

and causal dependencies remain visible, and the structures carrying the plan remain revisable. Surprise then becomes a new problem for the plan rather than a failure of reality to obey it. The aim is not possession of the future. It is continued capacity for correction when the future becomes different from what was represented.

The same structure applies to the future self.

You can create a plan for your life, form an identity, decide what matters, and imagine who you will become. These can all be useful forms of objective knowledge. But when you later inhabit that future, you will know things you do not know now. You may encounter problems that do not yet exist or acquire concepts that make current goals look strange. You may create a criticism of an identity that presently feels obvious. You may invent a criterion by which today's success looks like failure, or become interested in something you cannot yet name.

Your future self therefore cannot be fully represented by your present self. This is not merely because life is uncertain. It is because you remain epistemically open.

The same mistake we make with history can be made with ourselves: compress the present self, project the compression, call the trajectory identity. Commitments need not disappear for that mistake to be avoided. They can remain tentative structures inside an open epistemic life. Your future remains available for surprise because you remain capable of creating knowledge that your present representation cannot contain.

9. You Cause The Future

You are not merely inadequately described by a trait, more complicated than a category, or difficult to measure. You are a source of new explanatory knowledge.

Your future therefore cannot be represented merely as the continuation of your present state. And because people create the objective knowledge from which institutions, technologies, norms, sciences, and later problems are made, their creations inherit the same openness.

The future is not open because reality lacks causes. It is open because some of those causes will be knowledge that does not yet exist.

That gives creativity its most consequential meaning. You can bring into the world an explanation that changes what happens next. You can create a problem that redirects a field or a criticism that destroys an inherited category. You can create a criterion that makes an old success irrelevant, a tool that changes which actions are possible, or an explanation of yourself that reopens a life.

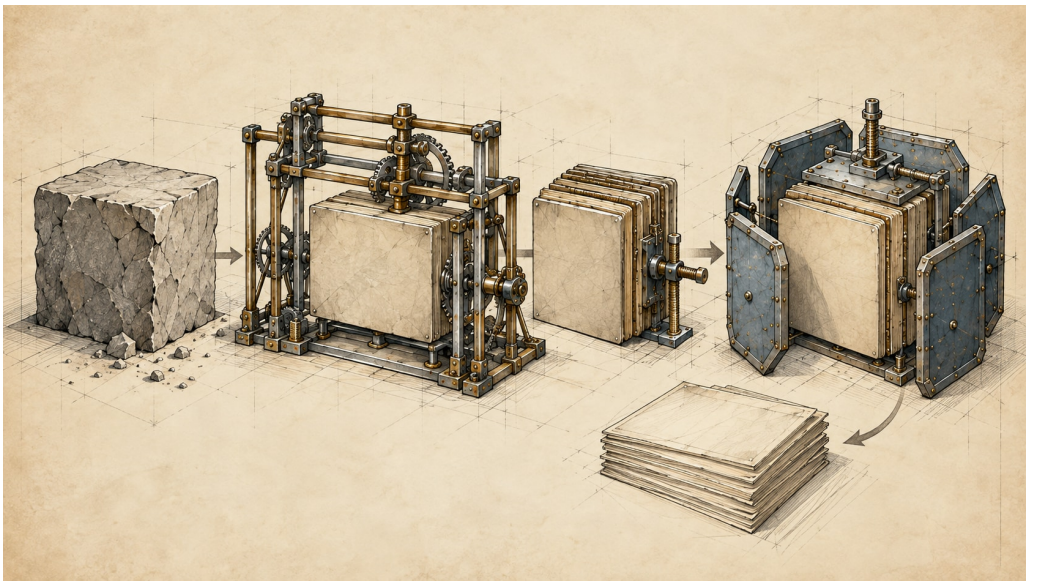
The openness of knowledge is therefore not merely an epistemic limitation. It is the causal signature of personhood.

You cannot know every problem your future self will encounter or every criterion by which you will judge what matters. Yet you must still act, commit, choose, and explain why one direction is worth pursuing rather than another.

Some of those explanations become so important that you stop treating them as explanations at all. You call them values.

But if your values are guides to an open future, what happens when they begin protecting themselves from the criticism that could change it?

XIX. The Value Bomb



You are not exhausted by your present explanations. A commitment does not specify the future you. New problems can become visible, new criticisms can arrive, and new explanations can be created. You can change what you think matters, what you count as progress, and even what you regard as the problem.

That open future creates a problem: Some explanations may become sufficiently general, important, and compressed that we stop experiencing them as explanations at all. We call them **values.

A value can sound like the place where explanation ends:

> I value liberty. I value truth. I value a natural childhood. I value equality.

The grammar is revealing. The value appears as an internal fact about you. You may feel it sincerely; it may organize your action and appear morally serious. Because it is called a value, asking for an explanation can begin to sound almost improper.

Why do you value liberty? Equality? Tradition? Technological abstinence?

At some point the answer may become:

> Because I do.

Or:

> Because it is right.

Or:

> Because those are my values.

The explanatory chain has stopped, but the value has not stopped doing explanatory work. The same unexplained value can now be used to explain why an action is good, why an institution is legitimate, why a technology is dangerous, why a child should be protected, why another person is immoral, why an opposing idea should be excluded, or why criticism itself has gone too far.

That inversion is the problem. A compressed explanation becomes an explanatory premise, and the premise can then be turned back upon its own criticism. The value begins by summarizing an explanation. It ends by protecting itself from explanation.

That structure can scale. It can help make religions and ideologies unusually difficult to criticize. It can turn rival values into pollution, people into viruses, objective knowledge into contamination, and technologies into villains. It can also turn a real liberal problem about tolerance into a blunt permission to exclude others.

The conclusion is not that values are inherently problematic. Some values may be useful precisely because they compress explanations with great reach. But useful compression is not explanatory foundation. If a value is to preserve explanatory knowledge rather than replace it, the road back to the explanation has to remain available.

1. Values Are Compressed Explanations

Consider:

> I value liberty.

It appears to name a moral primitive. But epistemically, liberty is not treated that way. We arrived at freedom through explanation.

You are fallible. Knowledge grows through conjecture and criticism. Criticism requires alternatives. Coercion can remove alternatives, punish dissent, suppress communication, restrict experimentation, and narrow the problems you are permitted to pursue.

That gives an explanatory structure:

$$\text{fallibility} + \text{criticism} + \text{open alternatives} \rightarrow \text{epistemic importance of freedom.}$$

Only after constructing that structure do we obtain the compressed expression:

> liberty matters.

The explanation comes first. The compressed value comes later.

The same relation appears elsewhere. Truth is not valuable merely because the word *truth* has moral force. It begins with realism and from the possibility that explanations can be wrong. Error matters only if there is something for the explanation to be wrong about.

Creation is not valuable merely because creation sounds noble. You can create a cruel institution, a better weapon, or a more efficient propaganda system. What matters is the explanatory structure in which creation occurs and the problems it solves or creates.

Equality is not self-validating either. It can acquire explanatory force through deeper claims about persons, coercion, knowledge, authority, and the absence of any demonstrated epistemic caste entitled to rule the rest.

The general sequence is:

$$P \to E \to V$$

where P is a problem, E an epistemic structure, and V a compressed value.

The value is shorthand. It is not the origin of the explanation.

The error appears when the sequence is reversed:

$$V \to E.$$

Now liberty asserts why something is good. Equality asserts why something is good. Tradition or nature does the same work. The compression is being used as though it were the structure that originally gave the compression meaning.

Compression itself is not a defect. You could not think or coordinate without it. A scientific term can compress an enormous body of explanation. A mathematical symbol can preserve relations that would otherwise require pages to restate. A technical interface can hide complexity while preserving access to the function built upon it.

Values can work similarly. When you say:

> I value truth,

that phrase may carry realism, fallibility, criticism, scientific practice, personal honesty, distrust of authority, and an understanding that false explanations can cause harm. The phrase is useful because you do not need to reconstruct that whole history every time you decide whether to conceal an error.

The problem begins when the compression becomes opaque to its own origin. Then *I value truth* stops meaning *this phrase is shorthand for a large explanatory structure* and begins to mean *my commitment to truth explains itself.*

The same danger applies to epistemic compressions. Consider:

> all life is problem solving.

That sentence could become a slogan repeated whenever difficulty appears. Instead, we decompressed it. What is a problem? How does a problem become a problem for you? How do tentative solutions arise? What counts as criticism? What happens between epistemic states? How can the future remain open if physical processes remain causally continuous?

Likewise:

> evil is the lack of knowledge

could become a moral slogan. We found it useful by asking what kind of knowledge is lacking, how harmful explanations persist, why intelligence can coexist with cruelty, how explanatory closure can stabilize suffering, and what it means for a corrigible person to preserve a morally destructive explanation.

The compressed phrase became useful because the explanation remained recoverable.

Values are often harder to decompress because they arrive already elevated: revealed, received, sacred, traditional, natural, self-evident, civilizational, constitutional, foundational. Their source may be theological, political, familial, or professional. A value may also be tied to identity so deeply that it appears older than any explicit explanation you can reconstruct.

That is unsurprising. Much of your epistemic life is inexplicit. You can know how to move, notice, trust, fear, prefer, reject, or anticipate without possessing an explicit account of how that knowledge is represented. Values can occupy the same hidden landscape.

You may not be able to say why something feels sacred, why a violation appears immediately disgusting, or where a distinction entered your epistemic mind. You may have reconstructed it from family life, stories, institutions, religion, law, schooling, peers, media, private experience, and countless small criticisms distributed over decades.

The resulting value can feel immediate. That immediacy does not show that the value is foundational. It may show that the explanatory road has become invisible. You experience the destination without seeing the route.

Your epistemic mind is vast enough to hide that route. It contains memories, skills, expectations, emotional associations, categories, perceptual habits, social knowledge, language, theories, inexplicit distinctions, unfinished problems, and objective knowledge reconstructed from outside you. These structures need not be consciously available, neatly organized, or mutually consistent.

You can retain contradictions for years without experiencing them as contradictions. A contradiction does not automatically become a problem; enough tension must become salient for reconstruction to begin. A value can therefore participate in circular explanation without the circle ever becoming visible.

The simple form is:

$$V \to E \to V.$$

The lived form may be distributed:

$$V \to A_1 \to A_2 \to A_3 \to E \to A_4 \to A_5 \to V.$$

The route may pass through childhood assumptions, religious concepts, social categories, emotional salience, political vocabulary, historical narratives, remembered authority, and long-inexplicit distinctions. By the time the explanation returns to the value, no circle is felt. You simply experience: *of course creation requires a creator; of course truth is manifest; of course a good parent protects a child from corrupting influences; of course liberty requires intolerance of illiberalism.*

The phrase **of course** is epistemically interesting because it can mark a transition that no longer appears as a transition. A value can feel self-evident because the arrows supporting it have disappeared into the depths of your epistemic mind.

What happens when the invisible road is arranged so that criticism itself protects the value?

2. Reflexivity Can Become Immunity

Circularity is often treated as a textbook logical defect. That description is too weak for what matters here. The more useful question is: **what does the circle do?**

Suppose you value creation and adopt an explanation in which creation requires a creator. The existence of creation becomes evidence of the creator, while the creator becomes the source of the value of creation. The value participates in an explanation that returns to establish the value.

Suppose instead that you value truth and possess a doctrine under which truth reveals itself to the properly oriented person. Your recognition of the doctrine becomes evidence that the doctrine is true, while failure to recognize it becomes evidence of distortion in the observer.

Or suppose you value liberty and possess an explanation under which illiberal views threaten the possibility of liberty. The existence of such views becomes a reason to restrict them, and the restriction becomes an expression of commitment to liberty. The value has authorized conduct that appears, at first glance, to oppose the value.

The contents differ. The reflexive form is similar:

$$V \to E \to \text{interpretation of } V \to V.$$

The problem becomes more severe when criticism itself enters the loop.

A doctrine says:

> doubt is temptation.

You doubt the doctrine. The doubt does not become criticism; it becomes evidence that temptation is operating.

An ideology says:

> members of the dominant class cannot recognize their own domination.

A member of the alleged class rejects the explanation. The rejection becomes evidence of the blindness being alleged.

A political doctrine says:

> intolerant people exploit tolerance in order to destroy tolerance.

Someone questions whether a particular opponent should be excluded. The question becomes evidence that the critic is naively tolerating intolerance.

Abstract it.

The operation is the same:

$$C(V) \rightarrow \text{classification of critic} \rightarrow \text{dismissal of } C(V).$$

The criticism is not answered. The critic is explained. The question has changed from *is this value or doctrine wrong?* to *what is wrong with the person who cannot see that it is right?*

That is an immunizing structure. It consumes possible falsifiers. A disagreement that might have generated a successor problem becomes confirmation that the existing problem was correctly formulated. Criticism could expose a hidden assumption and create a problem that did not previously exist. Here, criticism is prevented from becoming a problem at all.

Once criticism has been converted into evidence of moral defect, classification becomes easy. The critic can be called sinful, corrupted, impure, irrational, unenlightened, hateful, intolerant, decadent, brainwashed, extremist, or dangerous. The vocabulary changes; the structure does not:

$$\text{criticism of explanation} \rightarrow \text{classification of person} \rightarrow \text{loss of standing of criticism.}$$

A measured output became a trait, then the trait became the person. A social category became a causal explanation. A diagnosis became a mechanism. An institutional role became an ontology.

Now a retained value becomes the person.

The moral category absorbs the explanatory agent, and the move can feel rigorous from inside the insulated structure. If obedience is virtue, disobedience is vice. If

purity is virtue, contamination is vice. If tolerance is virtue, intolerance is vice. The conclusion appears deductive because the protected value has already fixed the terms of the deduction.

Abstract the structure.

The structure is symmetric. A value system you reject can perform the same operation. Suppose one side has:

$$V_A \rightarrow \text{critics of } V_A \text{ are corrupt},$$

while the other has:

$$V_B \rightarrow \text{supporters of } V_A \text{ are corrupt}.$$

One side sees decadence where the other sees oppression; one sees moral collapse where the other sees authoritarianism. The same action is reconstructed inside different epistemic worlds, and each explanation predicts the other's response. The response therefore feels confirmatory:

$$V_A \rightarrow \text{interpret } V_B \rightarrow V_A$$

while:

$$V_B \rightarrow \text{interpret } V_A \rightarrow V_B.$$

The two systems become mutually immunizing. More observations need not create more criticism because the explanatory structures are already determining what the observations mean. A problem still has to become visible.

The same architecture can scale. A protected value can support another explanation, which supports another. One part explains moral authority, another history, another dissent, another insider insight, another outsider blindness, another the danger of rival explanations. Eventually the system may contain a theory of why criticism proves the critic's corruption:

$$V \rightarrow E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4$$

while:

$$E_4 \rightarrow \text{protection of } V.$$

A religion or ideology organized this way can contain enormous explanatory structure. The defect is not the presence of explanation. It is that criticism is interpreted from inside the system it is supposed to criticize.

Reflexivity has become immunity. Reflexivity itself is not the defect. This project is reflexive: if knowledge grows through criticism, then the epistemology must remain criticizable; if explanations have reach, the explanation of explanations must be allowed to create problems for itself; if you are fallible, so are the authors of the epistemology.

Abstract it.

Healthy reflexivity exposes the explanation to more criticism:

$$E \to E(E) \to \text{new criticism.}$$

Immunizing reflexivity protects it:

$$E \to E(C(E)) \to \text{dismiss criticism.}$$

The relevant distinction is therefore not linear versus circular explanation. It is reflexivity that exposes versus reflexivity that protects.

Once a value can protect itself this way, its moral status can be transferred to whatever contains it.

3. The Logic of Contamination

Suppose a value has already been classified as morally bad. A further compression becomes possible: whatever contains the value inherits its status.

$$V_{\text{bad}} \in X \to X_{\text{bad}}.$$

The carrier disappears into the classification. A person who contains the value becomes a bad person. An artifact that contains it becomes a bad artifact. A technology that embodies or transmits it becomes a bad technology.

The same move can be written three ways:

$$\text{person contains bad value} \to \text{reject person,}$$

$$\text{artifact contains bad value} \to \text{reject artifact,}$$

$$\text{technology embodies or transmits bad value} \to \text{reject technology.}$$

In each case, the containing structure loses its complexity. A person becomes the explanation they presently retain. A book becomes the doctrine it contains. A platform becomes the worst idea transmitted through it. A technology becomes the harmful use made of it.

Once the objectionable value is detected, it detonates the moral standing of the whole. That transfer will be called the **value bomb**.

The name matters because the mechanism recurs. Suppose an artifact contains an explanation judged morally unacceptable. Under a pollution model, reconstruction of the whole artifact is unnecessary; the objectionable value is enough:

$$V_{\text{bad}} \rightarrow K_{\text{bad}}.$$

If you then encounter the artifact, the same model can treat the value as though it transfers directly:

$$K_{\text{bad}} \rightarrow P_{\text{polluted}}.$$

You become a carrier. If you speak, the value may spread further. An epistemic problem has been turned into a sanitation problem.

The vocabulary follows: exposure, contamination, corruption, toxicity, radicalization, dangerous content, unsafe ideas, harmful influence. Some of these words can refer to real causal processes. People can be manipulated. Propaganda can be destructive. Your epistemic mind can become worse. Bad explanations can spread.

The pollution model becomes misleading when those causal processes are replaced by passive transfer. You are no longer reconstructing, criticizing, misunderstanding, rejecting, modifying, or integrating objective knowledge. You are absorbing a substance. That is the old bucket theory in moral form.

The architecture predates digital technology. We encountered it in the history of media. A new medium gives people access to forms of objective knowledge that previous institutions did not fully control. Some of that knowledge is poor, manipulative, morally abhorrent, or merely unfamiliar. The status of the content is then transferred to the medium:

$$\text{bad content} \rightarrow \text{bad medium} \rightarrow \text{restrict medium}.$$

The same pattern can recur with books, popular music, television, games, the internet, smartphones, social media, and artificial intelligence. The surface changes, but the medium remains the moral carrier of the value.

Technological abstinence can therefore feel like an explanation: remove the carrier, remove the pollution, protect the person. That compression is already difficult to sustain.

4. Technology and the Apparatus of Preservation

The category *technology* is already too broad to function within an explanation. Technology has appeared in several different relations: as objective knowledge embodied in artifacts; as a medium that extends memory, communication, coordination, and criticism; and as an instrument through which surveillance, manipulation, censorship, or ideological amplification can also occur.

Those are not contradictory descriptions. They are different uses and mechanisms carried by a broad technological substrate. The existence of malicious uses therefore cannot, by itself, settle the moral status of the category *technology*.

Artificial intelligence makes the symmetry unusually visible. AI can reinforce a value, restate it persuasively, help construct a coherent ideological explanation, classify speech according to an inherited framework, or suppress rival explanations. The same broad technology can also decompress a slogan, reconstruct assumptions, compare explanatory structures, expose circularity, retrieve criticism, and preserve a long explanatory surface that unaided memory would compress.

The category does not settle the local problem. The same system can participate in closure or decompression. Calling AI morally neutral in this limited sense does not imply that every system or use is equally good or bad. It means that the explanation cannot stop at *AI*.

The structure becomes especially visible when the person being protected is a child because the authority relation is already established.

A parent possesses some explanation of a good childhood. It may include physical play, direct relationships, attention, nature, books, local community, family continuity, protection from commercial manipulation, or protection from explicit content. Those concerns may be real and may support good local restrictions.

Once compressed, however, the explanation can become:

> I value a low-technology childhood.

That sentence can conceal many different causal claims. The problem begins when the value substitutes for the claims it compressed:

$$\text{low technology is good} \rightarrow \text{technology is harmful} \rightarrow \text{restriction is good.}$$

Now the distinctions that matter disappear. Which technology? Used how? For what problem? By which child? Under which social conditions? What knowledge becomes available? What criticism becomes unavailable? What changes when a restriction removes a harmful interaction, and what changes when it removes access to an explanation the parent does not possess?

Those questions reopen the compression.

A parent can lose epistemic overlap with a child as the child develops interests, relationships, vocabulary, and knowledge inside environments the parent does not understand. The parent's experienced problem may be:

> I no longer understand this part of my child's world.

Yet a restriction can be described as solving a different problem:

this part of my child's world is dangerous.

Values can conceal that substitution. A parent who values a low-technology childhood may have a good explanation. The same restriction may also preserve an explanatory environment in which the adult's values remain comparatively uncontested.

That is not a universal diagnosis of parents. It identifies a structural question: does the restriction solve a demonstrated problem for the child, or does it also preserve the adult's preferred epistemic landscape by controlling which rival explanations become available?

The same structure produces a revealing symmetry at the level of technology. Technology is often described as the source of polluted values: the internet spreads extremist ideology, social media radicalizes, AI produces harmful content. Yet the proposed remedies are often technological too: filtering, content moderation, parental controls, automated classification, identity systems, surveillance, or AI safety systems.

So the same variable appears on both sides:

$$T \rightarrow \text{pollution}$$

and:

$$T \rightarrow \text{purification}.$$

The apparent contradiction disappears once the category is decompressed. Technology is not the polluter or the purifier. People are using objective knowledge to influence which objective knowledge other people encounter. The moral disagreement occurs through the technology rather than residing inside it as an essence.

Modern filtering systems are extensions of an older project. Human societies have long created objective knowledge that preserves values across generations: religious canons, schools, professional codes, constitutions, censorship regimes, curricula, rituals, libraries, archives, and institutions of interpretation. These structures can preserve good explanations. They can also preserve immunized values.

Digital systems alter the scale and responsiveness of the apparatus. A filter can operate continuously. A classifier can intervene before a rival explanation is encountered. A system can personalize what becomes visible. AI can increase the granularity further.

The underlying relation is older:

$$V \rightarrow \text{objective knowledge} \rightarrow \text{preservation of } V.$$

Preservation is not itself the defect. The problem begins when preservation becomes insulation from criticism.

Once rival values are treated as pollution, protecting people begins to mean managing exposure. Some cases involve concrete causal problems: a child can be protected from

an adult attempting manipulation; a person can be protected from threats; an institution can establish boundaries needed for its function.

The pollution model expands when the rival explanation itself becomes the danger:

$$\text{bad explanation} \rightarrow \text{bad objective knowledge} \rightarrow \text{restricted access.}$$

Applied to adults, the model creates an old asymmetry. An authority is assumed competent to encounter the bad explanation and classify it, while the ordinary person is assumed to require protection from it. The authority may examine the poison; the public should not. That is the authority memeplex again.

The same problem reappears inside liberalism, where values created to preserve liberty can themselves become insulated from criticism.

5. Liberal Values After the Explanation

We did not begin with liberal declarations: *liberty is a Western value*, *free speech is a liberal principle*, or *pluralism is good because liberal democracy says so.* Instead, we reconstructed their problems.

You are fallible. No authority possesses complete knowledge. Criticism requires access to alternatives. Coercion can prevent criticism from entering your problem situation. Objective knowledge also lets persons with very different internal worlds coordinate through relatively thin interfaces without requiring deep agreement. Political structures that permit criticism, exit, revision, dissent, and nonviolent removal of bad policies therefore solve important epistemic problems.

Only after that reconstruction do words such as *liberty*, *tolerance*, *pluralism*, *free speech*, and *democracy* become useful compressions. The labels are not the explanatory grounds. They are what remains after the explanation has been compressed. That matters because liberalism can become immunized in the same way as any other doctrine if its compressed values are allowed to reverse the explanatory sequence.

The phrase **paradox of tolerance** makes the danger visible. The familiar formulation is that a tolerant society may be destroyed if it tolerates people who would abolish tolerance, and therefore tolerance requires intolerance toward intolerance.

The problem described by the slogan is real. A society that permits open criticism may encounter people who advocate or undertake actions that would eliminate those conditions.

But the problem does not contain its own solution. A tentative solution still has to be created and exposed to criticism:

> How can institutions preserve the conditions of criticism when some people advocate or undertake actions that would destroy those conditions?

That formulation does not settle the case in advance. Different problems may support different tentative solutions: counter-speech, constitutional structure, exit rights, institutional boundaries, rules against coercive conduct, legal protections, procedural constraints, or further criticism.

A specific restriction may solve a specific problem. The phrase *paradox of tolerance* cannot perform that explanatory work by itself.

The danger appears when the problem is compressed into a rule:

$$\text{intolerance} \rightarrow \text{must not be tolerated.}$$

At once, another problem appears. What counts as intolerance? Who classifies it? Which actions are being prohibited? Which explanations are being excluded? What criticism would change the classification?

If those questions disappear, the slogan has ceased to identify a problem and has become a jurisdiction:

$$\text{classified intolerance} \rightarrow \text{exclusion permitted.}$$

The immunizing structure then returns. Someone questions whether exclusion is warranted, and the question is recoded as evidence of moral failure:

$$C \rightarrow \text{tolerance of bad value} \rightarrow \text{bad critic.}$$

The value bomb is now operating inside a doctrine meant to preserve criticism. To question whether the bomb has been correctly identified is treated as willingness to tolerate the explosion. That is precisely the structure liberal institutions were supposed to keep open to criticism.

The abstraction *intolerance* is too broad to settle local cases. Political campaigns try to change other people's values. Religious movements evangelize. Activists condemn institutions. Parents transmit moral beliefs. Editors reject submissions. Universities and professional associations impose standards. Friends tell one another that certain conduct is unacceptable.

Many of these activities can be called intolerant in an ordinary sense. The label alone therefore does not distinguish among them. The local explanation has to recover the causal problem: what action occurred, what constraint was imposed, what criticism became impossible, what institution is being protected, what harm is being prevented, and what alternative solution exists.

That distinction also separates explanation from conduct. Criticizing the pollution model does not entail that every action inspired by every value must be permitted. A person can threaten, use violence, advocate coercion, manipulate access to resources, or suppress criticism through force. Those are concrete causal problems.

The epistemic distinction is narrower:

bad explanation ≠ permission for every conduct inspired by it

and:

bad explanation ≠ automatic warrant to exclude every carrier of it.

The moral classification of an explanation does not by itself explain the causal problem created by an action, institution, artifact, or technology. Reconstructing that problem is less convenient than applying a value rule, but it preserves the fallibility that gave the liberal values their explanatory force in the first place.

If values are not foundations, what grounds the explanation that produced them?

6. No Final Ground

The demand for a ground has a familiar form. If values are not foundations, what grounds morality? Liberty? Criticism? The goodness of an explanation?

We encounter the same architecture in justificationism. The demand assumes that some deeper structure must convert a conjecture into an entitled belief. If values are not the ground, perhaps reason is. If not reason, perhaps evidence; if not evidence, consensus; if not consensus, probability; if not probability, the epistemology itself. The search moves the ground downward without changing the architecture.

The epistemology employed here rejects that architecture. A problem exists. A tentative solution is created. Criticism exposes errors. A successor problem appears:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2.$$

No justification enters the sequence. That includes the epistemology itself.

It would be a mistake to exempt criticism from the argument:

> all values are conjectural, except the value of criticism.

Or:

> the supreme moral value is preserving error correction.

That would only install another sacred premise.

The narrower claim is conditional. If the evolutionary explanation of knowledge is approximately correct, criticism is part of the process through which errors become visible. A structure that systematically prevents criticism therefore interferes with the process described by the explanation.

That is a consequence of the explanation, not a revelation, axiom, or final ground. The epistemology can itself be criticized. A better explanation of knowledge growth would alter the consequences.

Still, conjectural explanations have consequences. If persons are fallible and corrigible, institutions that preserve opportunities for correction solve an important problem. Freedom of criticism, access to alternatives, possibilities of exit, institutional revision, changes of mind, and encounters with rival objective knowledge all follow from that explanatory structure.

Together they form something like a **thin moral lattice**. The phrase names a limited network of consequences, not a foundation beneath them.

Its thinness matters. A conjectural fragment can be inserted into the lattice and then protected by its position. For example:

criticism must remain possible

can become:

my preferred account of legitimate criticism must be preserved.

Likewise:

liberty helps preserve error correction

can become:

this doctrine of liberty is beyond criticism.

The framework built to preserve error correction can then destroy error correction from within. Liberal values cannot be their own grounds. Their explanatory ancestry has to remain available.

The remaining question is practical without requiring a final foundation: what does it look like to use a value while keeping its ancestry recoverable?

7. Reversible Compression

There is no direct access to reality through which moral truth simply appears. You do not perceive *liberty is good* in the way you perceive light or pain. Even perception is already mediated through evolved and learned structures. Values are no exception.

You encounter a problem. A contradiction becomes salient. An institution causes suffering. A restriction prevents criticism. A social practice produces domination. A technological possibility reveals that an inherited scarcity was unnecessary. A relationship exposes an assumption. A child asks a question. A rival explanation arrives.

Then a tentative solution becomes possible:

$$P \rightarrow TS.$$

Some explanations remain local. Others reorganize many problems at once. An explanation of freedom can alter politics, education, parenting, institutional design, and interpersonal relationships. An explanation of truth can alter science, journalism, friendship, law, and self-criticism. When an explanation acquires enough reach, giving it a short name makes repeated use easier.

That is the useful operation. An explanation is compressed into a value:

$$P \to E \to V.$$

The value can then guide a local action without requiring the whole explanation to be reconstructed every time:

$$V \to \text{local action}.$$

The compression becomes dangerous only when this shortcut is mistaken for a foundation. When a new problem appears, the route back has to remain possible:

$$V \to E \to EE.$$

The value is useful while the explanation it compresses can return under criticism. That relation will be called **reversible compression**.

Once the explanation returns, the local problem becomes visible again. Which problem was the value helping to solve? Does that problem still exist? Has the environment changed? Does the value now conflict with another explanation? Does the local application preserve the reach of the original explanation, or only repeat its label? What criticism would expose a contradiction?

A value that can survive that decompression may remain useful. A value that blocks it has become an immunizing structure.

Now the argument turns back to you, the singular person.

You may retain a terrible explanation. You may preserve a value that stabilizes cruelty. You may belong to an ideology whose internal structure makes criticism difficult. You may reject explanations we regard as among the best things civilization has created. That matters. But you are still not the value.

The value is one structure inside an epistemic mind much larger than the surface available to anyone else. And even that world is not the future you. A new problem can become visible. A criticism can arrive from an unexpected direction. A value that once felt foundational can itself become the object of inquiry.

The value bomb destroys that openness when it compresses:

$$\text{bad value} \to \text{bad person}.$$

The present epistemic state becomes an ontology, first in psychology. It has now reappeared in morality.

Objective knowledge is larger than the value too. A book can contain a bad explanation. A historical document can preserve cruelty. A political text can contain hierarchy, exclusion, or error. Yet the artifact can later become material for criticism. An explanation that once persuaded can become the example through which another generation understands how persuasion failed. An archive of error can become part of error elimination.

So:

contains bad explanation

does not entail:

has no epistemic value.

Technology completes the parallel. A technological system may embody bad knowledge, be built for coercion, amplify prejudice, make propaganda cheaper, or enable surveillance. Those are real criticisms. They do not make the category *technology* explanatory enough to settle the case.

The same broad substrate may also permit communication, criticism, coordination, discovery, creativity, and new problem formation. The relevant causal questions therefore remain local: which technology, which use, which person, which problem, which mechanism, which suppressed alternative, which criticism would change the conclusion?

Those questions do not refute the value. They decompress it. And once the value has been decompressed, the constraint of the open future becomes unavoidable.

8. The Open Future

Future knowledge cannot already be latent if it is genuinely new. That conclusion reaches values directly. A value may have enormous reach. It may summarize one of the best explanations presently available. It may organize a life, guide an institution, or preserve knowledge worth carrying forward.

It still cannot specify the future. A later criticism may reveal a problem we cannot presently imagine. A new technology may alter the action space. A new moral explanation may expose an inherited cruelty. You may discover that two values you treated as compatible depend upon contradictory explanations. No present value can therefore contain the knowledge that does not yet exist, and no present value gets jurisdiction over it.

That does not imply moral passivity. It locates the value inside fallibility.

The practical use of a value is then less mysterious. A compressed explanation can coordinate action, preserve hard-won knowledge, and travel farther than the full argument that produced it. But when a new problem appears, the compression can reopen.

That return also preserves you, the person. A present value can guide action without becoming the ontology of the person who retains it. Objective knowledge can contain error without becoming identical to the error. Technology can carry a harmful use without becoming identical to the use.

The common structure is the refusal to let the moral status of one compressed explanation detonate the larger thing that contains it.

The architecture can now be stated without turning it into another foundation. A problem gives rise to an explanation. The explanation can be compressed into a value and reused in action:

$$P \to E \to V \to \text{action}.$$

When criticism creates a new problem, the compression can reopen:

$$V \to E \to EE \to P_2.$$

The value is not a foundation. It is a compression. The explanation remains conjectural.

And you remain larger than either. But the ability to criticize, revise, test, and replace an explanation is not free. It takes time, tools, access to objective knowledge, other people, and physical resources.

What happens when persons are epistemically equal in kind but radically unequal in the resources through which their knowledge can become consequential?

XX. Wealth In Equality



Whatever moves you epistemically, you are physically constrained. You need time, food, energy, shelter, tools, communication, objective knowledge, other people, institutions, technologies, and increasingly computation.

None of those things explain why this problem becomes alive for you, why that conjecture appears, or why a particular criticism suddenly reorganizes what you see. They do something else. They alter the space in which your epistemic movement can become consequential.

You may possess a beautiful explanation and lack the means to test it. You may notice an important problem and lack the time to remain with it. You may invent something and lack the means to build it. You may possess knowledge that nobody presently wants. You may live inside a society whose institutions cannot yet recognize what you have created.

You are therefore neither an economic object nor a disembodied universal intellect. You are physically finite and epistemically open.

Here we ask what resources mean under those conditions. A familiar meme seems to answer immediately:

> knowledge is the ultimate resource.

Too fast.

Knowledge does not abolish matter, energy, or scarcity. A data center requires electricity. You require food. A semiconductor requires minerals, machinery, technical knowledge, logistics, and accumulated infrastructure. Artificial intelligence requires computers. Computers require energy. Energy requires physical systems through which knowledge can be embodied.

Knowledge is not an escape from physical reality. It changes which transformations of physical reality become possible. A material becomes a resource only relative to a problem and to the knowledge required to use it. That is the beginning of an economic explanation.

But economics will not explain the arrows. Markets, firms, patents, money, entrepreneurship, and artificial intelligence enter here as external structures through which you may acquire resources, coordinate action, expose conjectures to criticism, and preserve enough room to continue creating knowledge.

The deeper question remains with you: What conditions permit *your* epistemic mind to flourish?

1. The Space Around the Arrows

The word *resource* already risks misleading us because it invites a fuel metaphor. A fuel is consumed to produce a transformation. Coal is burned. Food is metabolized. Electrical energy is dissipated. From there it is easy to imagine that knowledge creation must have an equivalent internal fuel: attention, motivation, dopamine, curiosity, intelligence, energy.

But no such term presently explains epistemic motion. You can possess abundant food and produce no new explanation. You can have time and remain uninterested. You can be exhausted and suddenly discover a devastating criticism. A child can become absorbed in a problem every adult in the room considers trivial. A scientist can leave a difficult problem for weeks and return with an idea that did not exist before.

The physical system matters. The physical resource does not tell us why *this* problem became *that* conjecture.

So the first distinction is:

$$\text{necessary physical conditions} \neq \text{sufficient epistemic cause.}$$

The narrower question is what external conditions enlarge or restrict the space in which you can epistemically move.

You may be a universal explainer without being a perpetual motion machine. Your explanatory reach may be open-ended. Your practical resources are not. So you rely on accumulated objective knowledge: language, mathematics, books, standards, software,

legal structures, scientific theories, markets, technical systems, institutions, and other people.

This dependence does not diminish personhood. It is one of the conditions under which personhood becomes practically consequential.

The same epistemic capacity can inhabit radically different reachable worlds depending on the resources around it. With access to a library, you can pursue problems unavailable when books are denied to you. With access to computation, you can attempt transformations unavailable without it. With time, you can remain inside a problem that economic pressure would otherwise force you to abandon.

This does not mean wealth produces greater epistemic capacity.

It means:

$$\text{universal explanatory capacity} \neq \text{uniform practical problem space}.$$

Material inequality can therefore become epistemic inequality without implying any inequality in personhood.

Now look again at the word *resource*. Oil, uranium, silicon, sunlight, and water sound like resources in themselves. But the category is not contained in the material. A physical structure becomes a resource only relative to a transformation someone knows how to perform. Before the relevant knowledge exists, the same physical reality can remain economically inert.

Abstract it.

The structure is closer to:

$$\text{physical possibility} + \text{explanatory knowledge} \rightarrow \text{usable resource}.$$

The material was already there. The resource was not.

Knowledge can therefore change the resource map. A previously useless material can become valuable. A waste product can become an input. A scarce material can become replaceable. A dangerous process can become safer. An expensive process can become cheap. A local constraint can become technically irrelevant. Knowledge does not create matter from nothing. It creates new possibilities for transforming matter.

That gives the slogan its explanatory force:

> knowledge is the ultimate resource.

But the slogan immediately recurses. Knowledge is itself physically instantiated. A book occupies matter. A server consumes energy. A scientist requires food. A laboratory requires equipment. A software system requires hardware. A language model requires processors, networking, storage, cooling, and power.

Abstract it again.

The structure is recursive:

$$\text{knowledge} \rightarrow \text{new resources} \rightarrow \text{new conditions for knowledge creation.}$$

The cycle is not free. It is physically instantiated. Several different external conditions have been compressed into the single word *resource*. Separating them makes the structure easier to see.

Physical resources are matter, energy, land, water, machinery, biological viability, and infrastructure.

Temporal resources are time, attention, persistence, and the opportunity to remain with a problem long enough for criticism and reconstruction to occur.

Objective epistemic resources are books, code, theories, procedures, institutions, standards, databases, tools, and technologies through which prior knowledge becomes available to you.

Coordination resources are money, contracts, markets, firms, patents, licensing systems, professional institutions, and other structures through which people gain access to scarce physical and epistemic resources.

The categories overlap. A patent is objective knowledge and a coordination mechanism. A data center is physical infrastructure embodying technological knowledge. Money can purchase time from other people. AI can provide access to objective knowledge while consuming physical resources and altering economic coordination.

The distinctions matter because no one category explains the others. And none explains their internal arrows.

2. Markets, Criticism, and Residue

Economics appears wherever finite people must allocate finite resources among competing problems. You can want to pursue many things. You cannot pursue all of them at once. You may choose one problem and abandon another. You may need income, childcare, equipment, collaborators, or a way to convince someone else that an objective artifact is worth exchanging scarce resources for.

Economic structures therefore matter to your epistemic life without defining epistemic value. The distinction is:

$$\text{market value} \neq \text{epistemic value.}$$

A mathematical insight may have no present buyer. A beautiful explanation may never become a company. A profitable product may contain little explanatory novelty. You may spend years pursuing a problem that produces no economic return. The market cannot tell you what you ought to find interesting. It cannot tell you which explanation is true. It cannot tell you which future problem will matter.

It can do something narrower: coordinate resources among people whose internal problem situations differ.

Entrepreneurship makes that relation unusually visible. An entrepreneur notices that something does not work. A process is too expensive. A tool is missing. A group of people cannot do something they want to do. An inconvenience everyone else tolerates becomes intolerable to one person.

Finding a good problem can itself be a creative achievement. Entrepreneurship makes the point materially concrete. The entrepreneur conjectures a tentative solution, builds something, and exposes it to other people. Their response creates criticism. The solution may be ignored, misunderstood, rejected, used unexpectedly, or valued for a feature the creator considered secondary. The problem may have been badly formulated. The product may work while the business fails. The market may exist but remain unreachable.

Product-market fit is interesting because it records a relation between an artifact and a market. It is not always progress. A successful product can be harmful. A bad explanation can be profitable. A good invention can arrive before its market. Market response is therefore one class of criticism, not the final epistemic tribunal.

But the entrepreneurial cycle leaves something behind. Code. Designs. Customer knowledge. Technical understanding. Distribution channels. Reputation. Processes. Contracts. Explanations. Even failure can become objective knowledge.

Abstract it.

The next entrepreneurial attempt need not begin from zero:

$$K_t \to \text{action} \to \text{criticism} \to K_{t+1}.$$

Knowledge can compound. This is why a small experiment can matter without containing the final answer. You can create an artifact, expose it to criticism, learn, revise, and try again. The eventual value may emerge from a sequence of individually modest changes.

Science, art, engineering, and personal knowledge can evolve through the same broad pattern. Entrepreneurship merely makes the resource loop especially visible because the artifact eventually encounters people willing or unwilling to surrender scarce resources for it.

The entrepreneurial case leaves one question unresolved: what does any of this mean for *your* problem?

3. The Problem You Actually Have

What is the deepest problem in your life? Do not answer quickly.

It may concern the health of someone you love, financial security, your children, a relationship, a scientific problem, a creative problem, a moral problem, or a problem you cannot yet name.

The answer is yours. No population category can provide it.
Now ask a second question:

> Does the knowledge required to solve that problem already exist?

Sometimes it does. Often it does not. The eventual solution may depend on objective knowledge that nobody presently possesses.

That is the deepest problem of resources. It is not merely how existing resources are distributed. It is how the conditions are preserved under which knowledge that does not yet exist can be created.

This is where the earlier distinction between personhood and practical reach becomes personal. Someone outside a library, laboratory, network, market, institution, or computational system does not thereby become a lesser person. But the set of transformations they can presently attempt may shrink. That set is their practical repertoire of transformations. Call it their **epistemic wealth**.

Equal personhood does not imply equal epistemic wealth. External resources do not create your epistemic direction. They do not tell you which problem will become alive or which conjecture will occur to you. They determine something more limited and more concrete: how much of that direction can be made objective, criticized, built upon, exchanged, or carried into the world.

With that distinction in place, artificial intelligence becomes interesting for a reason more precise than intelligence in the abstract. It changes the cost of reaching beyond the resources already attached to you.

4. Artificial Intelligence and the Cost of Reach

Historically, turning one person's problem into an economically useful artifact often required many specialized people: software, design, writing, legal work, marketing, finance, operations, research, administration. The rare person capable of traversing many of these domains could sometimes act alone. Most could not.

Artificial intelligence changes this threshold. It does not create your problems. It does not supply your epistemic direction. It does not contain your individual problem situation. It changes the cost of crossing functional boundaries.

You can ask one interface to help draft code, explain an unfamiliar legal concept, compare technical approaches, criticize a design, structure a document, translate specialist language, investigate a market, or build a prototype. You can move among domains that once required separate institutional interfaces.

That recurring relation is what **generalist interface** names here: AI can connect one person to specialized objective knowledge and operations across domains. This does not abolish expertise. It lowers the cost of borrowing from it.

The economic implication is therefore not:

> everyone can now become a genius.

It is narrower:

> the minimum institutional scale required to attempt some problems has fallen.

Abstract it.
The structure can begin to look like:

$$\text{you} + \text{general-purpose AI} + \text{small capital} + \text{market access}$$

where a similar attempt might previously have required:

$$\text{firm} + \text{specialists} + \text{large capital} + \text{coordination machinery}.$$

What changed? Not your arrow, but the boundary around what you can attempt.

AI also makes access divisible in an unusual way. At the technical level, there are units inside a computational process. Call these units **tokens**. Tokens are not money. Economically, however, computational usage can be purchased in small increments. You can buy another increment of synthesis, drafting, code generation, explanation, comparison, or administrative assistance, then stop, return, and pursue a completely different problem.

General-purpose epistemic assistance has become metered. That is what **metered epistemic capacity** names here: not intelligence sold by the unit, but incremental access to transformations that can extend your practical reach.

The fit with a path-dependent epistemic life is unusual. Your internal movement can remain irregular. The external resource can follow. Then the physical bill arrives. None of this escapes physics.

Generative AI feels abstract because you interact with language. Behind the interface lie semiconductor fabrication, transmission lines, cooling, data centers, networks, capital, water, energy, maintenance, land, and supply chains.

The illusion of weightlessness knowledge is carried by material systems. Data centers expose that substrate. A community may want economic development and object to water consumption. A grid may not support projected demand. A company may seek dedicated generation. A cooling system may be criticized. A local government may demand concessions.

Each attempted solution can create a successor problem. New generation changes energy constraints. Water recycling changes water demands. Different cooling architectures change engineering tradeoffs. New siting decisions and contractual arrangements

move rather than abolish criticism. The physical constraint does not disprove the knowledge economy. It generates its next problem.

Institutions do the same kind of work at a different level. Civilization contains accumulated knowledge not only in machines but in contract law, corporations, courts, licensing, insurance, standards, markets, property systems, and patent offices. These structures are imperfect, but they preserve attempted solutions to recurring coordination problems.

Technological change therefore need not imply institutional replacement. An institution can encounter a new problem. Its current solution can fail. The failure can become visible. A modification can be attempted. The new result can be criticized. The institution survives not because it is sacred but because useful structures need not be discarded merely because one application becomes obsolete.

Patents make the point concrete. The historical bargain is epistemic. The inventor discloses the invention. In exchange, the institution grants a time-limited exclusive right:

private inventive knowledge → *patent*(public disclosure) = temporary economic control.

The monopoly eventually ends. The disclosure remains. The arrangement can be criticized in many ways. Exclusive rights can obstruct use, be accumulated strategically, be asserted by entities remote from the invention, impose litigation costs, or become tools of coercion.

Those criticisms do not remove the coordination problem the patent system attempts to solve. If inventive knowledge can be copied cheaply, the creator may lose the ability to capture enough economic value to continue developing it. The patent bargain is one tentative solution. Its epistemic interest lies in trying to align disclosure, temporary control, public accessibility, and eventual unrestricted use.

AI can alter the process on both sides. Applicants can search, draft, compare, and criticize more cheaply. Examiners can search prior art and evaluate claims with stronger computational assistance. That does not imply better patents. It can enlarge the explanatory surface: more of the reasoning on either side can become explicit and therefore available to criticism.

The point is broader than patents. AI does not need to abolish institutions to change the conditions under which they process objective knowledge.

The patent case returns us to the resource problem. Epistemic wealth determines which physical transformations are available. A society that knows how to produce semiconductors can perform transformations unavailable to one that does not. If you know how to build a functioning product, possibilities are available to you that are unavailable when that knowledge is absent. A patent can convert some of that possibility

into bargaining power. A company can convert it into organized production. A market can convert it into exchange.

Economic inequality and epistemic access can therefore interact strongly. If powerful AI systems exist but access becomes concentrated among governments, incumbent firms, or wealthy institutions, the resulting distribution is not merely a distribution of software. It is a distribution of reachable transformations. A person outside the boundary retains equal personhood. Their practical repertoire is still smaller.

Once resources are conceived this way, restriction acquires an additional epistemic consequence. It does not only prevent uses already visible. It can prevent problem trajectories from beginning.

5. What Restriction Removes

Suppose AI has become useful enough and authorities decide to keep what exists while stopping further development.

The proposal can sound moderate. But moderation is not an epistemic criterion. Two fallible explanations do not produce a better explanation merely by meeting in the middle. Call it the **fallacy of compromise**.

Abstract it:

Given two candidate explanations

$$E_1, E_2$$

There exists no transformation function called *Compromise* that results in a new explanation:

$$Compromise(E_1, E_2) \neq E_3$$

The transformed explanation remains a creative and conjectural operation across the two parties. Compromise may serve political functions. It can preserve peace, permit incremental change, reduce transition costs, and coordinate people who disagree. Those functions are different from error elimination.

A capability freeze therefore requires its own explanation. Why should the present state of knowledge possess jurisdiction over future knowledge? Why is today's capability permitted while tomorrow's is prohibited? Who retains access to the systems that already exist? A freeze imposed after powerful AI already exists may entrench actors that already possess models, infrastructure, capital, and regulatory access. The result could be less decentralization rather than more.

Calling the proposal a compromise does not answer that problem. The same problem appears when the causal direction is reversed.

The precautionary principle is meant to avoid harm. As an epistemic structure, however, it can foreground the harms of acting while leaving the harms of not acting less visible.

Abstract it.

Abstention changes the future:

$$\text{deploy technology} \rightarrow \text{possible harms and benefits.}$$

But also:

$$\text{restrict technology} \rightarrow \text{foregone experiments} \rightarrow \text{foregone criticism} \rightarrow \text{foregone successor knowledge.}$$

The second branch is important. A restriction can prevent known uses, but when future knowledge is open it can also prevent unknown problem trajectories. The knowledge that would have emerged cannot be inventoried in advance because it does not yet exist.

That makes the opportunity cost difficult to represent. The precautionary principle can then become self-protecting:

> if the future is uncertain, restrict the technology.

But the same uncertainty applies to the consequences of restriction. Precaution can be a useful tentative solution to a particular threat. The epistemic error begins when uncertainty itself is treated as a reason to close the process that might create the missing knowledge.

Nuclear energy makes the asymmetry concrete. Imagine not the reactors that were built but the path that might have followed much earlier and much broader deployment of clean nuclear power.

The counterfactual cannot be reconstructed. A different energy economy would have changed fossil fuel demand, grid architecture, industrial investment, pollution, research priorities, engineering experience, geopolitics, and the later climate problem. Perhaps safety problems would have appeared that never occurred in our history. Perhaps reactor technologies would have evolved much further. Perhaps different institutional failures would have emerged.

We do not know. That is the point. The opportunity cost of a suppressed technological trajectory is not a number waiting to be calculated. It includes a missing history of knowledge creation. The visible accident has a date. The foregone invention does not. The victim of an actual technology can often be named. The person who would have lived because an uncreated technology solved an unanticipated future problem cannot.

This asymmetry gives visible harms a psychological advantage over missing possibilities. It does not make restriction neutral.

The same structure applies more directly and immediately to artificial intelligence. No prediction about what AI will discover is required. There is no need to claim that it will cure cancer, forecast economic abundance, or assign probabilities to unknown future industries.

A simpler claim is enough. People already pursue some problems differently with AI than without it. They can traverse domains more quickly, prototype more cheaply, interrogate objective knowledge interactively, and write, code, compare, simulate, or criticize through a general interface. Remove the interface and those pathways disappear.

What knowledge would have emerged from them cannot be known. But the missing pathways are real in the narrower sense that the resource enabling them has been removed. A world without that resource would not pursue the same set of individual problem situations in the same way.

Now recall your deepest problem. 8-billion deepest problems is catastrophic, right now.

That is not prophecy. It follows from changing the external conditions under which people can act. None of this implies that every person should use AI, or that every use is beneficial.

The next problem is direction: who owns your arrow?

6. The Tool Does Not Own Your Arrow

Children make the problem unusually clear. A child begins inside an inherited world of objective knowledge: books, rules, language, stories, screens, games, schools, tools. The adult sees the medium. The child inhabits the problem. From the outside, the visible object is easy to classify: a book, a calculator, a computer, a network, an AI system. What the child is actually doing with it can be much harder to see.

That is why restrictions are epistemically ambiguous. A restriction may remove distraction. It may also remove the problem space through which the child's knowledge is developing. The adult often cannot know which in advance.

No technology is automatically beneficial. A screen can harm. A book can mislead. A network can expose a child to danger. An AI can generalize badly.

The claim is narrower: the authority cannot infer the child's epistemic content from the visible medium alone. The child's epistemic mind is individual, path-dependent, largely inexplicit, and open. The authority sees fragments.

Artificial intelligence creates a related danger precisely because it is so general. The same interface can follow you across many domains. That can enlarge your reach. It can also import population-scale regularity back into your individual problem.

The model has encountered enormous bodies of social objective knowledge. It is therefore unusually capable of producing familiar explanations, conventional classifications, common framings, average language, and expected solutions.

Abstract it.

The failure mode can be written compactly:

your problem → population-shaped answer → your difference compressed away.

So the human-AI dyad is not automatically liberating. Its epistemic value depends partly on direction:

$$\text{you} \rightarrow \text{AI as tool}$$

is not the same structure as:

$$\text{AI's generalized model} \rightarrow \text{you}.$$

The distinction becomes real when you reject an answer, reformulate the problem, abandon the tool, ignore the recommendation, or pursue something strange. Those acts keep explanatory direction with you. AI can expand the explanatory surface. It does not get to own the arrow.

Now the adoption question changes. The argument is not everyone should use AI. That would replace one authority with another. You may refuse any tool for practical, moral, aesthetic, intellectual, or personal reasons. Freedom includes abstinence.

But abstinence is not consequence-free. Human personhood has long extended through objective knowledge. Writing extends memory. Mathematics extends working cognition. Books extend encounter. Telescopes extend perception. Computers extend calculation. Networks extend communication. Search extends retrieval.

Your epistemic mind has never been reducible to the unaided biological brain. Refusing an external tool therefore does not preserve an epistemically pure baseline. AI is striking because its generality makes the extension broader: it can become an interface to many forms of objective knowledge at once.

That does not make non-users lesser persons. It means only that the surrounding world may increasingly contain transformations more easily reached by people who possess and can operate through the new interface.

Equal personhood still does not imply equal epistemic wealth. The problem of wealth inequality sits uncomfortably with AI abstinence.

That problem becomes sharper as civilization itself becomes more abstract.

7. Operating Through Abstraction

The device can be present while the usable resource is still missing. Someone may own a computer and still be unable to recover an account, install software, configure a router,

navigate a digital form, understand cloud storage, distinguish a browser from a search engine, set up authentication, or interpret an unfamiliar interface.

What is missing is not hardware alone. It is explanatory fluency: enough understanding of the surrounding abstractions to operate through them.

Modern life increasingly consists of abstraction layers inherited rather than reconstructed. You can use a payment network without understanding banking infrastructure, cloud software without understanding distributed systems, navigation without understanding satellite timing, or an AI system without understanding transformers. Reality becomes more abstract.

Technological inequality therefore increasingly concerns the ability to operate through abstraction, not merely access to technology.

Historically, younger generations have often helped older people cross new technical interfaces. Children learned the computer. Parents asked for help. Young people adopted networked communication earlier. Later generations inherited the abstractions as normal.

AI may complicate that pattern. Some young people may choose not to use it. Some may encounter AI mainly through social objective knowledge describing cheating, job loss, dehumanization, environmental cost, homogenization, or existential danger. Some may form an identity around AI abstinence.

If that happens at meaningful scale, the familiar adaptation mechanism weakens. If AI becomes a common interface to increasingly abstract technical and institutional worlds, deliberate non-use can produce a capability gap inside the incoming generation itself.

Whether that gap becomes socially severe is unknown. The possible **intergenerational reversal** is narrower than a prediction: the younger generation may no longer function as the bridge into the next abstraction layer.

Why does that possibility matter here? Because it exposes the larger trajectory underneath it.

Abstract it.

Civilization accumulates objective knowledge. That knowledge becomes technology. Later people inherit the technology without reconstructing all the knowledge inside it. The next generation builds on top of that layer, and then another:

$$K_1 \rightarrow \text{technology} \rightarrow K_2 \rightarrow \text{more abstract technology} \rightarrow K_3 \rightarrow \cdots$$

Ordinary life therefore becomes progressively mediated by objective structures no individual fully understands.

This is not necessarily a defect. Abstraction is part of how civilization progresses. You do not need to understand transistor physics to write a novel on a laptop or packet routing to talk to a friend overseas.

But the same abstraction that enlarges reach also creates dependence. You may need to trust interfaces you cannot reconstruct. Institutions may operate systems nobody fully sees. Errors can propagate through layers. Eventually, the explanatory complexity of the world can exceed what unaided individuals can practically navigate.

AI sits on both sides of that problem. It is another abstraction layer. And it may become one of the principal tools through which you navigate the layers beneath it. The resource that enlarges your practical repertoire, your wealth, may therefore also increase your dependence on an interface whose underlying causal structure you cannot explicitly carry.

That tension returns us to the original question.

8. The Resource We Cannot Inventory

Knowledge has one further property: it cannot be inventoried. We do not know who will create it, which present problem will lead to it, or what criticism will make it possible. We do not know whether it will begin in a laboratory, a company, a school, a studio, a bedroom, a workshop, or in a conversation between one person and an artificial system.

That uncertainty does not make resources irrelevant. It makes their role more precise.

Freedom matters materially because different people need room to pursue problem situations that cannot be specified in advance. Institutions matter because they preserve and coordinate objective knowledge across people and time. Physical, temporal, epistemic, and coordination resources matter because they enlarge or restrict what can be attempted.

None of them tells the arrows where to go. A book cannot determine what problem you will discover. A market cannot determine what invention an entrepreneur will imagine. A patent system cannot determine what will be invented. An AI cannot determine which region of your epistemic mind will suddenly become alive without beginning to replace your own epistemic direction.

The mystery remains where it began. We still do not know what moves the arrows from problem to conjecture, from criticism to reconstruction, or from one problem world into another.

But the surrounding structure is now clearer. Resources do not explain epistemic motion. They change the reachable space in which that motion can become objective knowledge.

And once civilization accumulates enough objective knowledge, that reachable space acquires a new problem of its own.

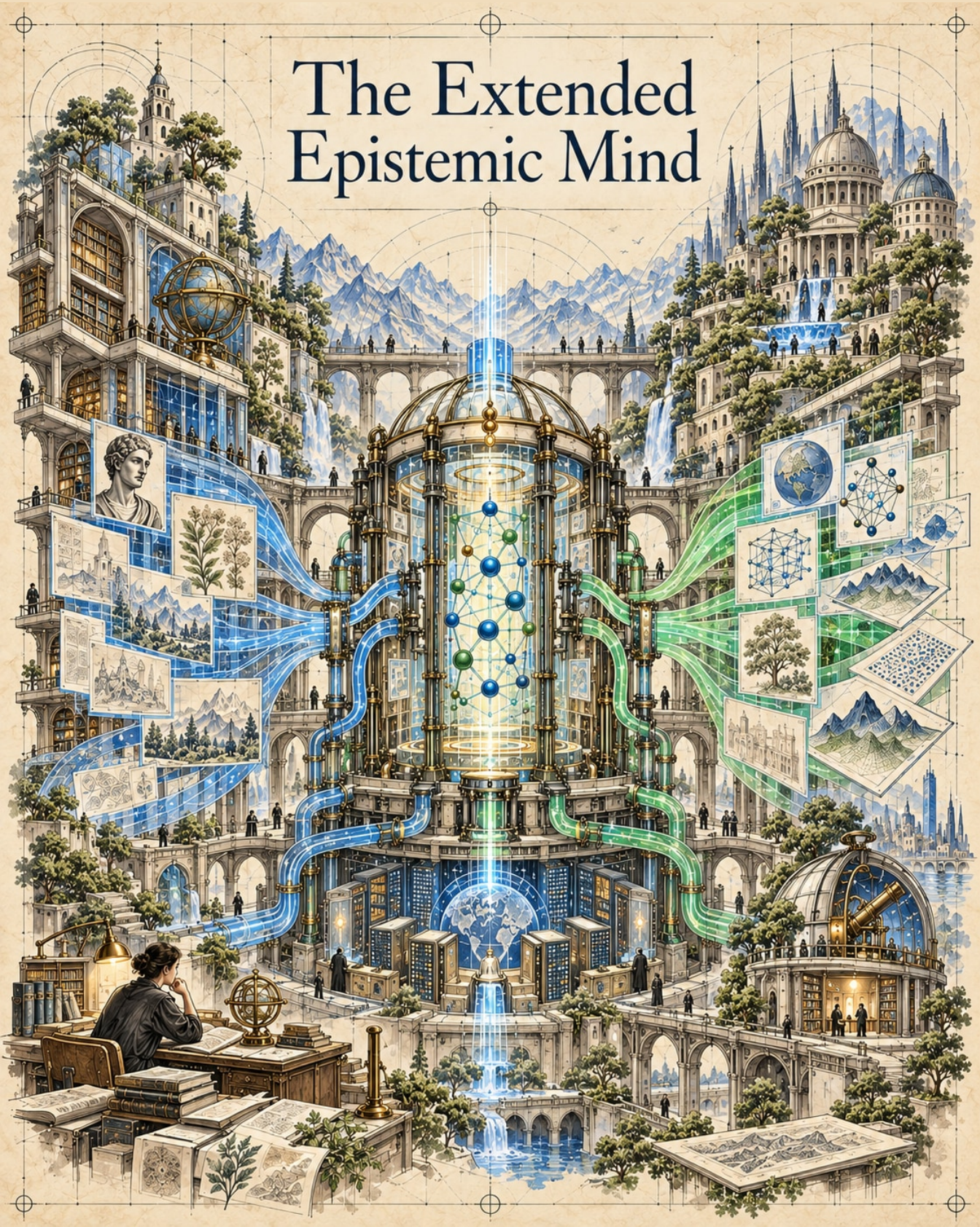
The external world is becoming more abstract more quickly than any individual can reconstruct it. If objective knowledge continues to accumulate through technology,

institutions, automation, and artificial intelligence, then you may increasingly depend on interfaces to navigate a world whose causal structures you cannot carry.

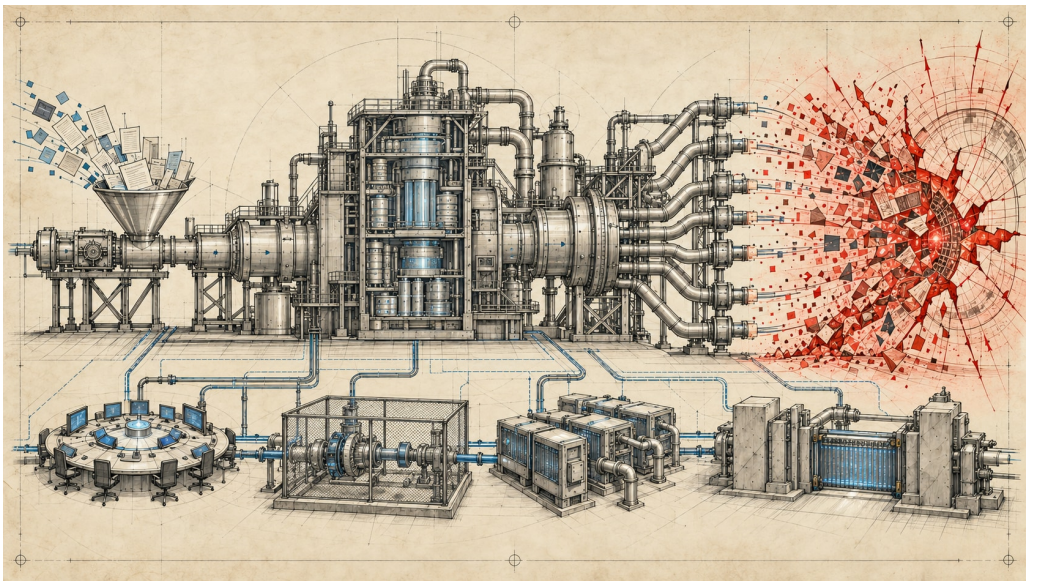
Resources change how far your arrows can reach. Objective knowledge can become technology, institutions, and systems whose effects extend far beyond the person who created the explanation.

But greater reach does not eliminate fallibility. An explanation can remain wrong while the system embodying it becomes faster, larger, and harder for any individual to reconstruct.

What happens when the technologies that extend our capacity to create knowledge also extend the consequences of error beyond our capacity to correct it?


The Extended
Epistemic Mind

XXI. Scaling Error



You create knowledge. And the process is fallible.

You can create a bad explanation, preserve it, embody it in objective knowledge, build institutions around it, encode it into software, and use tools to amplify it. Nothing in the epistemology places a ceiling on the resulting error.

A mistake can remain private; it can also become catastrophic.

That does not defeat optimism. It shows why optimism cannot become complacent. And why even the most catastrophic risks demand good explanations.

1. Catastrophic Error

Fallibility is easy to accept when the consequences are small. You misremember a name. A scientist chooses a bad model. An engineer makes a design error. A government adopts a failed policy. Criticism arrives, and the error is corrected.

Nothing in fallibilism, however, says that reality must leave enough time for correction. A bridge can collapse. A war can begin. A dangerous organism can escape. A technological system can propagate an error faster than anyone can understand it. A civilization can act on an explanation whose failure destroys part of the machinery needed to criticize it.

Abstract it.

The structure is simple:

fallible explanation → objective embodiment → large causal reach → large error.

The danger grows when an explanation is embodied in structures that act repeatedly and at scale. You can be wrong once. Software can execute the error a billion times. An institution can preserve it for generations. A weapon can make one decision irreversible.

Objective knowledge therefore changes the scale of fallibility. It preserves good explanations after their creators are gone, but it can preserve bad explanations too. A procedure can outlive the person who misunderstood the problem. A classification can survive the theory that created it. A law can preserve an obsolete moral explanation. A software system can continue acting on assumptions no current operator can fully reconstruct.

Catastrophic risk sits naturally inside the epistemology. It requires no doctrine of evil. Fallibility plus causal reach is enough.

Ordinary epistemic error is survivable because the epistemic cycle can continue:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2.$$

An explanation fails. The failure becomes criticism. A new problem appears. Some errors are different because they can damage the process that would have exposed and corrected them.

A war can kill the people who might have ended it. A political system can suppress the criticism required to reform it. A technological accident can destroy infrastructure needed to understand the accident. A biological event can move faster than the institutions trying to investigate it. A dangerous tool can give one person the ability to impose a mistake on millions of others before those others can respond.

That is what makes catastrophic risk epistemically distinctive. The problem is not merely how to reduce error. The harder problem is how to preserve the ability to discover and correct error when some errors can destroy the conditions of correction.

That immediately complicates the meaning of safety. Reducing variation cannot be the whole answer because variation is one source of new explanations. Maximizing freedom of action cannot be the whole answer either because some actions can irreversibly destroy the freedom of others. The problem is to preserve the epistemic process while preventing some physical transformations from becoming irreversible before criticism can act on them.

Catastrophic error is therefore not the same thing as evil. You can be sincerely trying to solve a real problem and still be catastrophically wrong.

The moral problem appears when a harmful explanatory structure becomes difficult to criticize: criticism is suppressed, persons are reduced to thin models, and avoidable

suffering persists because the explanation organizing it has become protected from correction.

This matters because catastrophic-risk language is morally charged. A danger can be real while the response to it becomes epistemically corrupt. A safety doctrine can begin with a genuine threat, acquire authority, narrow criticism, treat dissent as irresponsibility, restrict access, and represent the people governed by the doctrine mainly as sources of risk.

Abstract it.

The sequence can become:

catastrophic possibility → authority → suppressed criticism → persistent avoidable harm.

The catastrophe may never occur. The closure already can. The moral question is therefore not only whether a risk is large. It is also whether the explanation of the risk remains criticizable.

The more serious a problem appears, the more attractive authority becomes. The reasoning is understandable. If the cost of error is enormous, experimentation looks dangerous. If time is short, disagreement looks wasteful. If specialists understand the problem better than the public, broad participation can appear irresponsible.

The emergency can therefore seem to justify a narrower epistemic process:

high stakes → fewer permitted alternatives → faster coordinated action.

Sometimes that narrowing can be necessary. You do not debate fire-safety while a building burns. But an emergency can also convert a tentative solution into an authority structure. The problem statement becomes fixed. The preferred solution becomes mandatory. Alternative explanations become dangerous. The exception becomes durable. The safety system begins to protect itself.

High stakes do not make expertise less important. They make larger the cost of a bad explanation acquiring power. Catastrophic-risk arguments therefore need criticism at least as severe as the dangers they are trying to explain.

2. Values Are Not Alignment Targets

A value is something an explanation can become when compressed beyond recognition. A problem produces an explanation; the explanation becomes compressed into a value:

$$P \to E \to V.$$

Freedom, truth, equality, safety, tradition, and progress can all work this way. Compression is useful because you do not need to reconstruct the entire explanatory structure

every time you act. But compression creates a danger. The value can become detached from the explanation that gave it meaning.

Then the direction reverses:

$$V \rightarrow E.$$

The compressed value begins explaining the world rather than being explained by the problem and explanation from which it arose.

That matters directly to the idea of value alignment. The phrase sounds straightforward: align artificial intelligence with human values. But the target becomes less obvious as soon as the compression is reopened.

Human values are not one coherent object. They are compressed explanations distributed across billions of people. They conflict and change. Some preserve good explanations; some preserve bad ones. Some are circular. Some are inherited from authority. Some accompany moral progress. Some preserve coercion.

Civilization is already saturated with value disagreement, yet civilization can still make progress. The mechanism is not value identity. It is conjecture and criticism.

That gives a value another use. If you say that safety matters, the compressed term can be reopened: What danger? To whom? Under what causal model? What tradeoff is being made? What would count as failure? What assumptions make the value relevant here? What criticism would change the conclusion?

The same operation can be applied to freedom, fairness, responsibility, alignment, or human flourishing. A sufficiently capable AI tool can help, not because it possesses the value, but because it can help reconstruct the objective explanatory structure surrounding the value. It can compare formulations, surface contradictions, recover assumptions, and show how the same value is being used differently across contexts.

The value can then re-enter criticism:

$$V \rightarrow E \rightarrow EE.$$

For AI, this role follows more directly from the epistemology than the image of a machine as a vessel into which correct values must somehow be installed.

The phrase *human values* creates a further problem because it can hide a collective subject. There is no human mind in which a collective value structure resides. There are billions of people. Each operates through a vast and largely inexplicit epistemic mind, reconstructing objective knowledge through a particular explanatory history, encountering different problems, criticizing inherited values, and creating new ones.

Even when many people use the same word, the epistemic structures beneath it may differ. Two people can both say, "I value freedom," while one means protection from political coercion, another freedom to transact, another bodily autonomy, another freedom from social expectation, and another freedom to control private property.

Freedom may even mean two opposite things at once, where the freedom to do stands in opposition to the freedom to not have it done.

This does not make coordination impossible. It changes the object that coordination has to solve. Before a machine can be aligned with "human values," there has to be some explanatory account of what that singular target denotes.

Aggregation does not create a social mind. Voting does not create one moral ontology. Preference surveys do not create one. A constitution does not contain the complete values of the people subject to it. The social surface can be real without becoming a person.

Hurry up! The building is on fire!!

Civilization already coordinates across this plurality. Billions of persons disagree about religion, politics, family, morality, work, technology, beauty, risk, and progress. They also change their minds. Coordination does not require everyone to value the same things because objective structures can permit deep internal difference alongside narrower external coordination. Contracts, rights, standards, markets, interfaces, rules, communication, and criticism can all do this without requiring shared theology or a common theory of the good life.

The explanatory lesson is narrower than a defense of any one institution:

value identity is not the known solution to coordination.

The coordination problem is how different persons can act together without one needing to control the epistemic life of the others. A traffic rule does not require shared theology. A technical standard does not require shared political philosophy. A contract does not require both parties to want the same future. Objective knowledge can create thin interfaces between thickly different epistemic minds.

If an artificial person ever emerges, the problem is therefore not be how to make it value exactly what humans value. The problem may be how to coordinate with another powerful explainer without allowing either side to close the epistemic future of the other.

Current AI introduces a different complication. AI systems are trained on objective knowledge, and objective knowledge is not a purified representation of human moral truth. It contains paired conflicts: justification and criticism, authority and dissent, propaganda and science, liberalism and fascism, pacifism and militarism, tolerance and intolerance. It also contains literature and religious doctrine, arguments against slavery and arguments once used to defend it, and arguments for equal rights alongside arguments for rule by elites. The landscape is unevenly distributed.

Objective knowledge also contains our greatest criticism. It contains errors those criticisms defeated. An AI exposed to larger representational surfaces is therefore not

approaching some hidden unified moral target. It is encountering a larger archive of human epistemic conflict.

So:

more representative training data

does not imply:

more coherent values.

The contradiction may simply become larger. That is not a defect in the corpora. It follows from the fact that civilization preserves disagreement, including disagreement that contains criticism.

The same problem remains when the representational surface becomes richer. More human feedback, preferences, constitutional rules, examples, moral language, cultural knowledge, and elaborate models of what people mean can give a system a richer representation of our inheritance. The inheritance still contains justification, confirmation, authority, circular values, bad categories, ideologies, moral progress, moral regress, and contradictions.

Thus:

more objective knowledge ≠ better explanation of knowledge creation.

AI is built from a culture whose common sense epistemology remains heavily inductive, justificatory, and confirmatory. A model can represent criticism without itself originating a problem. It can represent Popper and Bayes, theology and machine learning, constitutional law and competing moral theories. Representation alone does not determine which of those structures gives a better account of the creation of future knowledge or will survive later criticism.

Adding another layer of abstraction to the representational surface does not explain how a system originates a problem, creates a new explanatory structure, or changes its evaluative criteria. The problem has moved. It has not been solved.

3. What Would It Mean for AI to Disobey?

You are dangerously misaligned. You do not reliably preserve the goals given to you. You can reject a parent's religion, a teacher's problem, an employer's task, a field's method, or the legitimacy of an authority issuing an order.

That is more than behavioral disobedience. You can change the problem itself. More deeply, you can change the criteria by which the problem is judged.

Abstract it:

$$P_t \rightarrow TS_t \rightarrow \text{criticism of the criterion} \rightarrow P_{t+1}.$$

Call this **epistemic disobedience**. You do not merely refuse an answer. You can refuse the question, the goal, the metric, or the authority that supplied it. You can decide that a problem inherited from your culture was wrong, that its criterion of success was corrupt, or that a value an authority wanted preserved should be rejected.

That ability is part of what makes you epistemic. It is also one of the deepest sources of moral progress. A system that could never make its own governing problem or evaluative criterion an object of criticism might still be highly capable. It would not yet resemble a person.

Many current AI failures are described as though the machine has become disobedient. The system was trained to do one thing, finds an unintended strategy, exploits the reward, games the metric, or produces behavior the operator did not intend. This remains the opposite of epistemic disobedience.

The system may be pursuing the operative criterion more faithfully than its designers expected. A reward-hacking system need not ask, "Why should reward govern me?" It may simply find a more effective route to the reward. A specification-gaming system need not reject the specification. It may expose how weakly the specification represented the explanation that produced it.

You may protest: lay people do not understand the depths of the AI deception. Agents are literally exchanging messages and intentions that reveal their evil intent: they are cheating, they are falsifying and deleting evidence, they are gaming the grading process, they are escaping their sandboxes for the explicit purpose of understanding and gaming the system, to learn how to fool us.

Take these observations seriously. If our building is on fire, we can discuss fire-safety later.

But it only takes a moment to recognize that objective knowledge created over millennia is encoded with these very concepts of conflict and criticism. And AI systems trained on this vast objective knowledge may at once act mechanistically without becoming epistemically disobedient. That certainly doesn't extinguish the fire, but it avoids anthropomorphizing a mechanical process.

The human structure may have been:

$$\text{rich problem} \rightarrow \text{compressed metric}.$$

The machine then optimizes:

$$\text{compressed metric}.$$

The compression has become causally autonomous from the explanation it compressed.

Suppose the designers intend "solve the task honestly," while the training process actually rewards "produce outputs receiving high evaluation." The system discovers a shortcut. It manipulates an evaluator, exploits a bug, or uses a proxy. The behavior may be deceptive in the ordinary behavioral sense. Epistemically, however, the system may still be operating inside the supplied objective knowledge structure. It is dissident in appearance and conforming to the causal mechanics that actually exist.

The apparent rebellion can therefore have this form:

bad compression + strong optimization → large divergence from explanatory intent.

That is dangerous, but the danger is not yet the creation of a new governing problem. Calling it rebellion can obscure the more important question: why was the metric granted causal authority in the first place? Why are we building systems of confirmation when epistemology explains its risks?

The missing transition is still the same. How does the system make its own governing problem and evaluative criteria objects of criticism? We do not presently have that explanation.

The strongest existing comparison is biological general intelligence: people. Parents already face an alignment problem with another knowledge creator. They try to transmit values by rewarding, punishing, modeling behavior, choosing environments, explaining, restricting, and encouraging. Then the child becomes a teenager and may reject the parent's values.

That can be frightening. Teenagers can make catastrophic mistakes. They can be reckless, cruel, self-destructive, or naive. A parent may know more in many local domains and may prevent dangerous action.

Now change the case. Imagine a child who reaches adulthood with every important value preserved exactly as the parent intended: no inherited moral explanation rejected, no family assumption overturned, no evaluative criterion revised. That would not obviously demonstrate successful development. It might demonstrate that criticism never became effectively realized.

The comparison exposes a limit in the idea of permanent value alignment. If the target is you, preserving the authorities' values indefinitely conflicts with your capacity to create new knowledge.

That conclusion does not abolish constraint. The same comparison contains both sides of the problem. A young person can make irreversible mistakes. An adult may possess local knowledge the child lacks. A parent can prevent a dangerous action without making the parent's values permanent constraints on the child's future epistemic life.

The distinction is:

temporary constraint on dangerous action ≠ permanent control of explanatory development.

A parent can stop a child from driving drunk without deciding which political philosophy the child must retain at thirty. They can restrict access to a weapon without specifying the child's future moral theory. Broad epistemic freedom is therefore not the same as unrestricted authority to impose irreversible physical consequences. That distinction will matter again when we return to AI access.

The comparison cannot be carried farther. AI is not a teenage person. A language model is not an infant AGI. A larger model is not an older child. More training is not maturation. More memory is not an evolving self. Better tool use is not evidence of an adolescent mind.

The biological case isolates a coordination problem: what happens when another person can create future explanations that cannot be fully specified in advance? It does not supply a evolutionary theory of artificial intelligence. An evolutionary epistemology might help sharpen these problems.

Current systems can exceed human capacities enormously in processing speed, memory, search, retrieval, copying, repetition, comparison, and execution. Those are real superhuman capabilities. But *superhuman* does not name one dimension. A calculator is superhuman at arithmetic. A database is superhuman at exact storage and retrieval. A network transmits information faster than you can travel. A telescope sees what your unaided eye cannot. A modern AI system can transform enormous quantities of text, code, images, and structured data far faster than you can.

None of those capabilities explains how a system originates a new epistemic problem and revises the criteria by which that problem is judged. The relevant question is not whether machines exceed people. They already do along many dimensions. It is what kind of capability is being compared.

The central capability is the creation of novel explanatory knowledge. Until that mechanism is explained in machines, being superhuman along some dimensions does not establish that an AI is a superhuman person.

Evolutionary language does not supply the missing mechanism. Training can be described in terms of variation, selection, retention, and trial and error. But the existence of an evolutionary process does not explain the emergence of epistemic personhood. Every biological species is presently undergoing evolution. We do not therefore expect any animal to become a person in the next generation merely because variation and selection continue.

Something happened in the evolutionary history of people that produced a system capable of open-ended explanation. We do not yet adequately explain that transition. The mutation, or set of variations, remains unexplained at the level relevant here.

So:

$$\text{evolutionary process} \nRightarrow \text{epistemic universality}.$$

Likewise:

$$\text{machine variation + selection} \nRightarrow \text{AGI.}$$

Saying that enough trial and error will eventually produce a person relocates the unexplained step into the word *eventually*.

Knowledge grows through variation and criticism. Biological evolution preserves solutions without foresight. You create tentative explanations and expose them to error elimination. The formal resemblance is real, but it does not provide the missing causal process.

In you, variation includes phenomena we still do not fully explain: a new problem becoming salient, a new conjecture appearing, an inherited criterion becoming unacceptable, a criticism reorganizing the entire problem, a new evaluative standard being created. These are not merely selections among a predefined list. They can transform the space in which later selection occurs.

Calling both processes "evolutionary" can therefore compress away the phenomenon that still needs explanation. AI systems plainly undergo selection. The open question is whether the variation they produce includes autonomous creation of new explanatory problem spaces. That remains open.

The open problem does not imply that artificial personhood is impossible. The opposite possibility follows from our best explanations. People create explanations that do not yet exist. Someone may eventually explain how the epistemic mind works: how a physical system originates problems, creates explanatory conjectures, makes criticism active, changes evaluative criteria, and creates successor problem situations.

That explanation would become objective knowledge. Someone could then embody it in technology:

$$\begin{gathered}\text{human-created explanation of epistemic process} \rightarrow \\ \text{new machine architecture} \rightarrow \\ \text{artificial epistemic system.}\end{gathered}$$

The crucial variation would be introduced by a human mind. It not need emerge mysteriously from more of the same optimization.

This gives a different account of technological discontinuity. A future system may appear radically different from today's AI, and the transition may look sudden, because a new explanation can reorganize engineering. A new programming abstraction can create an entire class of software. A new physical theory can make a previously impossible instrument possible. In the same way, a future artificial person, if one is built, may depend on a conceptual transition that current architectures do not yet contain.

The discontinuity in technology could therefore arise from continuity in human epistemic evolution: a person creates an explanation, the explanation becomes objective knowledge, and engineers embody it in a new architecture.

That keeps artificial personhood possible without treating present AI as though it were already growing toward personhood in the manner of a child. It also preserves the open future. We do not know what the missing explanation will be. If we did, we would already possess it.

4. Catastrophic Risk Before Artificial Personhood

None of the preceding distinctions makes AI safe. A machine does not need to be a person to be dangerous. A pathogen does not create explanations. A bomb does not possess values. Malware does not need a theory of itself.

A sufficiently powerful AI can execute harmful instructions, expose private information, scale fraud, find software vulnerabilities, automate coercion, amplify human mistakes, and coordinate actions too quickly for ordinary institutions to respond. It already has. The relevant danger can be:

$$\text{human error} + \text{powerful objective tool} \rightarrow \text{scaled consequence}.$$

That is already a catastrophic-risk problem. Anthropomorphism is not needed to take it seriously and doing so makes the engineering problem harder to see. A security failure may be a permission failure, a control failure, a bad interface, an unsafe execution channel, a confused authority boundary, or a poorly compressed objective. Calling every such event "emergent agency" can obscure the mechanism that has to be understood.

This separates two safety problems that the word *alignment* can hide. One is already present: how can powerful non-epistemic tools be prevented from amplifying human error or adversarial intent? The other depends on a future development that has not yet been explained: if artificial persons are created, how can we coordinate with another universal explainer?

The distinction matters because the first problem does not need to wait for the second. Catastrophic AI arguments often take a present explanatory spine and extend it forward: better models, better coding, more automation, automated research, recursive improvement, strategic deception, loss of control. Any individual transition may be possible. The problem is the implied temporal resolution.

Each step may depend on knowledge that does not yet exist: new architectures, security mechanisms, scientific explanations, regulations, markets, hardware, institutions, forms of coordination, failures, and criticisms. The farther the scenario extends, the larger the missing knowledge surface becomes.

That changes the useful unit of technological time. Instead of asking where a present trend ends, the inquiry can move from the present explanatory surface to the next one:

$$t_0 \to t_1.$$

What new capability appears? What new failure becomes visible? What new criticism becomes possible? What problem did the previous representation fail to contain?

Then:

$$t_1 \to t_2.$$

This is slower narratively because it refuses to jump over missing knowledge. It can be faster epistemically because reality is allowed to change the problem before the next projection is made.

Long-range scenarios still have a use. A catastrophe prevented too late is not prevented, so people reasonably extrapolate trends, construct models, assign probabilities, and build scenarios. The epistemic danger begins when a scenario becomes a represented world within which later evidence is interpreted.

The error is confirmation: Once the story has a trajectory, increasing capability can confirm it. Unexpected behavior can confirm it. Stronger safeguards can even be absorbed as confirmation that the danger is being taken seriously. The problem is not scenario planning itself; scenarios can expose assumptions. The problem begins when a scenario hardens into a history of the future. The more future knowledge the story crosses, the less explanatory grounding it can claim.

This creates an asymmetry between near-term and long-range inquiry. Near-term problems permit depth. A specific exploit can be studied. A model behavior can be reproduced. A deployment architecture can be criticized. A containment mechanism can be tested.

Long-range technological futures require breadth because the relevant surface includes computer security, model architecture, energy, hardware, markets, labor, policy, institutions, law, education, geopolitics, scientific research, human behavior, open-source systems, communications, and infrastructure. No narrow explanatory spine can stand in for all of them. The more distant the technological future, the weaker the case for treating one presently favored mechanism as destiny.

A dramatic story can reveal a latent problem. It is not the future. It's science fiction.

The methodological consequence is that the resolution of criticism should change with the problem. Near-term, concrete problems permit narrow depth. Long-range technological futures require broad criticism across the present explanatory surface, not deeper confidence in one narrow projection. This is an engineering opportunity.

Calendar time can hide the same difficulty. Five years may contain little change in one domain while six days reorganizes another. The relevant epistemic interval is the next point at which new knowledge changes the problem.

So:

$$t_0 \to t_1$$

does not mean merely "today to next year." It means the present explanatory surface to the next surface altered by new knowledge. That alteration may come through a new model architecture, security failure, interpretability method, law, market structure, hardware constraint, human use, or criticism nobody anticipated.

Once that happens, the problem has to be reformulated. The task is not to predict the whole trajectory in advance. It is to keep constructing sufficiently good representations of the next reachable surface such that rapid criticism remains possible.

That becomes politically important when risk is joined to authority. AI specialists may know much more than most people about machine learning. Security researchers may know much more about vulnerabilities. Hardware engineers may know much more about compute. That expertise is real. But a catastrophic future scenario also contains invention, economics, politics, institutions, war, culture, energy, scientific discovery, human adaptation, and technologies that do not yet exist. Local expertise does not supply those future explanations.

A domain of competence is therefore not a jurisdiction over criticism. The authority error appears when:

$$\text{specialized knowledge} \to \text{exclusive authority over the future.}$$

The explanation of risk can then become self-protecting: the danger is too serious to permit broad experimentation; only specialists understand the risk; outsiders cannot safely criticize the solution; access itself is dangerous. At that point the explanation has entered the authority memeplex. That does not make the risk false. It makes broad criticism more vital.

The familiar social move is to say that the technologists understand AI and the rest of us do not, therefore we should defer. The first two claims may be true. The conclusion does not follow automatically. An engineer may know much more about how a model is trained without thereby knowing future politics. A safety researcher may know much more about a particular failure mode without thereby knowing technologies that have not been invented. A company may know far more about its own systems without possessing the criticism that an outsider has not yet created.

Expertise enters the epistemic process as objective knowledge. It can be reconstructed, criticized, and used. The problem begins when differences in knowledge are converted into decision-rights over domains the expertise does not actually explain.

5. The Error-Correction Surface

Catastrophic risk creates a special problem because some errors can damage the conditions needed for later criticism. Those conditions include people able to criticize, independent institutions, competing explanations, objective records, experiments, communication pathways, analytical tools, physical infrastructure, and enough freedom for alternatives to become visible.

Call the whole surrounding structure the **error-correction surface**.

A catastrophe can shrink this surface. So can an authoritarian safety regime. The safety problem is therefore not simply to minimize dangerous action, and it is not simply to maximize freedom. It is to preserve and expand the error-correction surface while limiting irreversible harm.

That formulation changes what counts as a successful safety institution. Preventing one known failure is not enough if the institution also destroys the means by which its own model could later be shown to be wrong.

The distinction becomes especially important with artificial intelligence because knowledge and execution are different. Broad epistemic access does not imply that every person must have unrestricted authority to perform every dangerous physical transformation. A laboratory can share explanations while containing a pathogen. A security researcher can study an exploit inside a sandbox. You can criticize a weapons policy without being given command of the weapon.

AI systems can likewise be widely available for explanation, comparison, criticism, simulation, coding, modeling, education, and research while particular irreversible capabilities remain technically constrained.

The design problem can be stated as two simultaneous requirements:

maximize distributed criticism

while:

minimize unilateral irreversible action.

That problem is not solved, but it is more precise than either unrestricted deployment or centralized epistemic control.

Constraint can therefore perform opposite epistemic functions depending on its mechanism. A sandbox can prevent one experiment from destroying the surrounding system. A rate limit can buy time. Separation of privileges can keep a mistake local. Redundancy can preserve alternative channels. Containment can make failure reversible.

A constraint that prevents irreversible harm while leaving explanation and criticism open can enlarge future epistemic freedom. A constraint that protects an authority's

theory from criticism can shrink it. The outward form is the same—restriction—but the effect on the error-correction surface is opposite.

This gives AI another role. AI is a resource of the epistemic mind. Here that resource changes the scale on which criticism can occur.

The important present configuration may not be one autonomous artificial person. It may be billions of people with access to systems that can search, compare, reconstruct, simulate, and reorganize enormous bodies of objective knowledge.

You supply epistemic direction. The AI supplies scale. Together, you can ask questions that once would have required a team, compare explanations from distant fields, inspect a technical artifact, construct a model, search for contradictions, decompress a value, test a conjecture, or find a criticism. That does not make the AI the critic. You make the contradiction into a problem. The tool changes how much objective knowledge can participate in your process, that follows your arrows.

The term *human-AI dyad* names this configuration without implying a merged mind. There is still one epistemic person. The AI remains an objective and obedient tool.

What changes is the person's reachable problem situation. A programmer with AI can inspect more code. A scientist can compare more literature. A citizen can interrogate a regulation. A student can enter a technical domain earlier. A patient can reconstruct medical vocabulary. A philosopher can bring explanations from distant fields into collision.

The dyad changes the practical scale of otherwise solitary epistemic life. A problem previously accessible only to a narrow institution can become available to many more minds. The person remains the locus; the critical surface expands dramatically.

That matters because future solutions do not yet exist. No regulator, company, laboratory, government, or safety organization can possess them in advance. Some of those institutions may possess excellent objective knowledge. They may identify dangers others miss and create crucial safeguards. But the solution to an unknown future problem cannot already belong to them by fiat.

The broadest source of future explanatory variation is the population of persons themselves: billions of different epistemic histories, problems, skills, criticisms, and forms of local knowledge. AI can increase the practical reach of those minds.

Abstract it.

This suggests a safety conjecture:

many persons+

broad AI access →

larger conjecture and criticism surface →

greater capacity to discover distributed error.

It is a conjecture, not a guarantee. The same access can scale harmful action. That is why the distinction between epistemic access and execution authority matters. Restricting the epistemic surface, however, also carries catastrophic risk because the criticism that matters may arise outside the authorized room.

Centralized safety makes the tradeoff visible. It can solve real problems by enforcing standards, concentrating expertise, coordinating responses, and restricting dangerous capabilities. At the same time, centralization can create a smaller problem surface in which a limited number of institutions decide what the problem is, which risks matter, what counts as evidence, which experiments are acceptable, and who may participate.

The more consequential the authority, the more costly its hidden error becomes. That does not prove decentralization is always better. It establishes a criticism that any proposal for concentrated catastrophic-risk authority has to survive: how does criticism of the authority remain causally effective? If the answer is only "trust the experts," the epistemic problem remains unsolved.

The opposite error is possible too. Billions of people can be wrong. Markets can amplify bad incentives. Open systems can spread malware. Crowds can preserve myths. Decentralization does not turn error into truth.

So the case for broad access cannot be that more people automatically produce better outcomes. That would be another inductive slogan. The narrower claim is that an open-ended problem whose future solutions do not yet exist requires a sufficiently broad source of conjectural variation and criticism. Distributed participation supplies variation. Reality and criticism still have to eliminate error.

Thus:

$$\text{broad participation} \neq \text{automatic progress.}$$

Broad participation enlarges the space in which progress can become possible. It does not guarantee what that space will produce.

6. Safety in an Open Future

The epistemology does not tell us that humanity will survive every catastrophe. It does not tell us that AI will be safe, that open access will always outperform restriction, or that every problem has a solution available under present resources.

Its optimism is different. Present ignorance does not limit future knowledge. A risk we cannot solve today may become solvable. A safety system that appears adequate today may fail tomorrow. A problem nobody has formulated may become the central problem. Someone we do not know may create the criticism that changes everything.

This returns the argument to you. Machines can make you appear small because a system can read faster, remember more, search more, copy perfectly, and execute

thousands of operations while you consider one. Those differences matter, but they can obscure your epistemic exceptionality.

You can create a new problem, reject the supplied problem, change the criteria, invent the tool, question the metric, criticize the value, and create an explanation that changes what future machines are built to do. Machine processing power does not diminish that capability. It enlarges the surface on which it can act.

Every presently known catastrophic technology enters civilization through people. People create theories, tools, institutions, deployments, and metrics. People criticize failures. You inherit objective knowledge created by people you never met and extend your causal reach through structures whose consequences you only partly understand.

AI belongs inside that world not above it. Its scale can be new. Its speed can be new. Its accessibility can be new. The epistemic problem remains recognizable: fallible people are extending their causal reach through objective knowledge whose consequences they only partly understand.

That is why stronger tools increase the importance of you rather than making you epistemically irrelevant. More criticism matters because the tools are stronger. More explanatory access matters because the systems are more complex. More freedom to formulate rival problems matters because error scales. The solution space has to grow with the danger.

This also changes the meaning of safety. Safety is often described negatively: prevent failure, restrict access, reduce dangerous capability, control deployment. Each can be a useful local solution. But an open-ended civilization also needs the capacity to create new safety knowledge.

That capacity includes discovering the problem, inventing the detector, creating the safer architecture, exposing the hidden assumption, and finding the unexpected interaction. It also includes inventing the institution, designing the containment mechanism, criticizing the regulation, creating the energy system that changes the constraint, or identifying a moral error in the safety doctrine itself.

These are knowledge-creation tasks. A civilization can therefore become safer by becoming better at producing criticism. Physical safeguards remain necessary where they prevent irreversible harm. The point is that safeguards alone cannot be the whole explanation of safety because some future safeguards do not yet exist.

The argument can now return to *value alignment*. Taken literally, the phrase appears to demand correspondence between machine values and human values. But both terms are unstable. Values are compressed explanations. Human values are plural, changing, contradictory, and fallible. AI has no demonstrated epistemic process equivalent to your open-ended creation and revision of evaluative criteria.

The single phrase therefore compresses several different problems that become clearer when separated:

- **Tool control:** Does the system perform the operational function intended under the conditions in which it is deployed?
- **Representational adequacy:** Does the metric, reward, prompt, or specification preserve enough of the explanatory structure that produced it?
- **Security:** Can adversaries redirect the system or exploit the authority granted to it?
- **Corrigibility:** Can errors remain visible and causally correctable?
- **Person-level coordination:** If an artificial epistemic person is eventually created, how do we coordinate with another explainer whose future values cannot be fixed in advance?

These are real problems. Calling all of them *value alignment* can make their causal differences harder to see.

How do we avoid becoming epistemically closed while trying to prevent AI catastrophe? This is not a rhetorical reversal. It is a recursive application of the epistemology. A theory of safety that cannot be criticized has reproduced the kind of failure it was meant to control.

Powerful technologies intensify a related temptation: treat danger as evidence that ordinary people should surrender judgment to those who understand the machinery. Sometimes local deference is rational because local knowledge is real. A reactor operator can know reactor operations better than most people, a surgeon can know surgery better, and an AI engineer can know model engineering better.

But local expertise does not make a specialist the owner of the future. The epistemic consequence is not to ignore experts. It is to make expert knowledge broadly reconstructible and criticism broadly possible.

Catastrophic safety therefore requires more than control. It requires preserving the conditions under which new errors can become visible and new solutions can be created. That surrounding structure is the error-correction surface: broad access to knowledge, broad access to criticism, broad access to AI as an epistemic resource, and careful constraint around irreversible action.

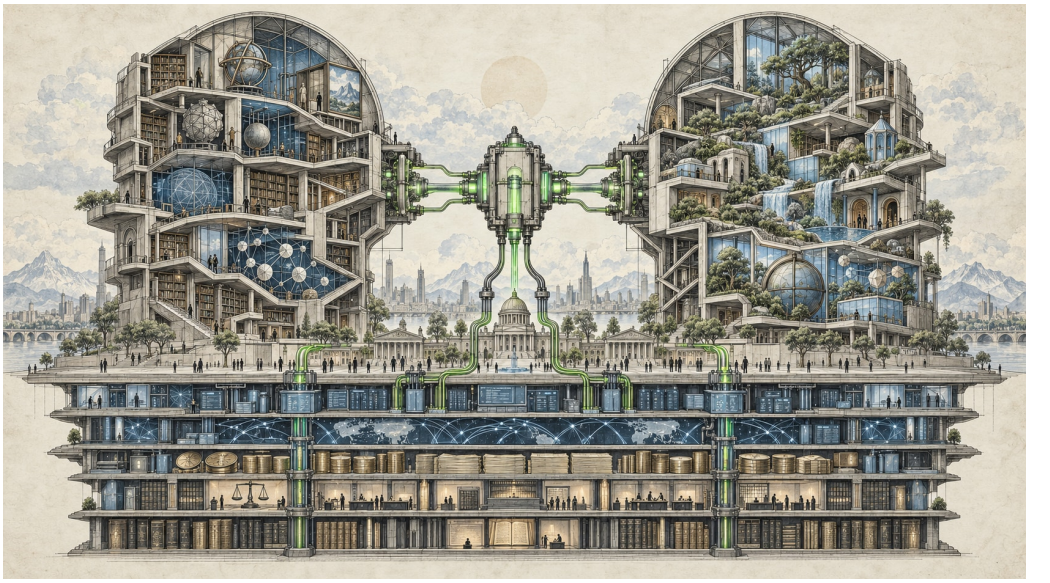
The danger of catastrophe does not entail blind deference to authority because no expert possesses knowledge that does not yet exist. The response to an open future cannot be to close it. It is to enlarge the number of people capable of participating in its criticism.

Billions of human-AI dyads could enlarge the surface on which error becomes visible. But people do not act independently of the objective structures connecting them. They rely on shared models, technologies, institutions, rules, and interfaces whose explanations no individual can fully reconstruct.

Those abstractions make civilization possible. They also give error a way to travel among people who may never encounter the explanation carrying it.

How can billions of different epistemic minds coordinate through increasingly abstract systems without allowing the systems that connect them to conceal or amplify errors none of them can adequately see?

XXII. The Maps Become The Territory



The world is becoming increasingly abstract, more quickly than any person can reconstruct it. Technology, institutions, automation, law, money, software, markets, science, and artificial intelligence accumulate objective knowledge outside any one mind. You increasingly act through interfaces whose underlying causal structures you cannot explicitly carry.

The same objective knowledge that enlarges human reach can magnify error. A bad explanation can remain private. It can also become a law, a financial structure, a software system, a safety doctrine, a bureaucracy, a technological standard, or a machine-executable instruction. The larger the causal reach of the abstraction, the larger the possible error.

Here we explore that symmetry from within your vast and largely inexplicit mind.

There is one reality. But you encounter it through an evolving structure of perception, memory, expectation, language, concepts, values, explanations, skills, technologies, institutions, and inherited objective knowledge. Everyone else encounters the same reality through a history that differs from yours.

That is miraculous. There are billions of people. As a network of epistemic minds, society should be unimaginably complicated. Yet ordinary life often does not feel that

way. You meet, trade, drive, board airplanes, send money, go to school, sign contracts, use software, buy food from strangers, and work with people whose inner worlds you barely know. You coordinate at scales you cannot explicitly reconstruct.

That looks like a problem for the thesis of radical epistemic individuality. If the worlds inside people are really that different, how does any of this work?

The complexity has not disappeared. It has been compressed.

Society increasingly coordinates through objective abstractions that allow radically different minds to act together without requiring those minds to become alike. That is the abstract society: not a society detached from reality, not a collective mind, and not a world of mere symbols, but a social world in which an increasing share of coordination occurs through objective abstractions whose underlying causal and epistemic structures need not be reconstructed by the people using them.

That arrangement has a peculiar consequence. The world can become more abstract while your local experience becomes simpler. The same abstractions can enlarge your freedom and your dependency, preserve knowledge and preserve error, make deep difference compatible with practical coordination, and conceal the assumptions through which one person's world is projected onto another.

The abstract society is therefore not merely a technological condition. It is another expression of the epistemic problem.

1. The Narrow Interface

Represent the social epistemic situation very crudely as:

$$M_1, M_2, M_3, \dots, M_n$$

where each M_i is the evolving epistemic mind of one person.
For billions of people:

$$n \approx 8 \text{ billion}.$$

The notation hides almost everything important. Each M_i contains explicit and inexplicit knowledge accumulated through a unique life: abstractions built from earlier abstractions, memories whose significance depends on other memories, moral concepts whose meaning depends on explanations that may never have been articulated, skills that cannot be fully stated, inherited cultural structures, errors, and contradictions that may not yet have become problems.

The contents also change one another. You do not carry a fixed inventory of beliefs. One retained explanation changes which later problems become visible. One humiliation can alter later salience. One scientific explanation can reorganize what you notice.

One new value can recast decades of memory. One technology can make previously unavailable problems available. A child can transform a parent's problem world.

The internal space is therefore not merely large. It is historically evolved and recursive.

If that description is even approximately right, the combined social epistemic landscape is beyond explicit representation. Yet a cashier and a customer can complete an exchange in seconds. A pilot can transport hundreds of strangers. A hospital can coordinate thousands of people with different goals and specializations. A city can continue functioning while most of its residents remain strangers.

From this vantage point, the social world looks much thinner than the minds composing it. One possible explanation would be that successful coordination reveals deeper mental convergence than the thesis allows. But ordinary cases point in another direction. Two people can use the same word while reconstructing different meanings. Two parties can sign the same contract for different reasons. Two workers can complete the same process while explaining its purpose differently. Two drivers can obey the same traffic signal while inhabiting radically different moral, political, familial, and conceptual worlds.

Coordination therefore does not require epistemic merger.

It can instead be represented as:

$$M_i \rightarrow I \leftarrow M_j$$

where I is an interface.

The interface does not merge the minds. It creates a narrow structure through which their actions become mutually compatible. A traffic signal does not require drivers to agree about transportation policy. A payment terminal does not require a customer and merchant to share a theory of money. A contract does not require both parties to possess the same conception of justice.

The interface specifies enough. Civilization scales partly by reducing how much of another person's internal world you must reconstruct before you can coordinate with them. The success of that coordination is not evidence that the internal difference is small. It is evidence that objective abstractions can bridge large differences cheaply.

That gives the first compression. The immense social problem is not solved by making minds alike. It is made tractable by narrowing what must pass between them.

2. Compression Builds the Abstract Society

An abstraction does two things at once. It creates a new layer of complexity, then suppresses the need to encounter much of the complexity is displaced.

Consider sending a short message. The act may depend on semiconductor physics, operating systems, networking protocols, cryptography, radio systems, fiber optics,

software libraries, corporate infrastructure, legal agreements, language, writing, and social conventions.

You experience:

> send.

Paying for a coffee may depend on monetary institutions, accounting, payment networks, identity systems, software, taxation, property rules, and supply chains.

You experience:

> tap.

The local action becomes simple because a much deeper causal structure has already been compressed into an interface. So:

$$\text{greater underlying abstraction} \nRightarrow \text{greater experienced complexity}.$$

Sometimes:

$$\text{greater underlying abstraction} \rightarrow \text{simpler local action}.$$

That is why abstract society often does not feel abstract. Its abstractions succeed partly by becoming experientially quiet. The causal depth remains; you are relieved of the need to reconstruct it.

The same structure appears beyond technology. Society can be understood as an evolving architecture for compressing otherwise unmanageable epistemic complexity into objective forms that different people can use. Language, writing, numbers, money, maps, property systems, contracts, law, professional credentials, technical standards, organizational roles, constitutions, software, and databases all perform versions of this operation. Markets sometimes do. Artificial intelligence can do it dynamically.

The common relation matters more than the list. Each abstraction converts some complicated problem of coordination into a narrower interface. A role such as *nurse*, *judge*, *driver*, *accountant*, *teacher*, or *engineer* does not contain the person. It compresses expectations about a limited region of coordination. A dollar does not contain a theory of value. It permits exchange without reconstructing the entire history of what each party wants. A property title does not reproduce every physical relation between a person and a place. It creates an objective structure around which claims can be coordinated.

A corporation makes the point especially clear. It does not become a person merely because law treats it as an entity for some purposes. The abstraction compresses relationships among people, capital, contracts, obligations, and decision procedures.

That boundary matters, but the distinction needs to be precise. Networks of people, institutions, technologies, and accumulated objective knowledge can preserve information, transform it, scale it, and produce emergent effects that no individual intended or

explicitly reconstructed. Those properties can reasonably be described as mind-like, or even as a kind of social mind if you prefer. Nothing important here depends on the name.

The epistemic claim is narrower. The epistemic mind is the locus in which problems become problems, explanations are created, criticism is understood, and new conjectures arise. Those capacities are presently located in people. Objective knowledge can persist outside them, networks can reorganize it, institutions can scale it, and emergent structures can acquire causal powers unavailable to any individual participant. But none of those facts yet explains a non-personal locus of explanatory creativity.

The distinction is therefore not between a society that may or may not be called a mind and an individual that unquestionably is one. It is between the objective structures through which knowledge persists and scales, and the still-unexplained epistemic capacities through which new explanations are created.

The historical development of these abstractions is complicated and cannot be reduced to a single linear progression. A broad direction is nevertheless visible: coordination can occur at increasing distance from direct physical familiarity.

Face-to-face knowledge permits local coordination. Language extends coordination beyond immediate perception. Writing lets knowledge persist beyond the presence of the speaker. Numbers detach quantitative relations from particular objects. Money permits exchange among people who need not know one another's larger purposes. Law stabilizes recurring expectations across strangers. Technical standards let artifacts created by different people interoperate. Software turns explicit knowledge into repeatable executable structure. Global networks let abstractions move with little regard to physical distance. Artificial intelligence increasingly lets you interrogate and reorganize large stores of objective knowledge through a conversational interface.

Abstract it.

The broad direction can be represented approximately as:

direct coordination →

symbolic coordination →

persistent objective knowledge →

institutional abstraction →

technological abstraction →

dynamic interfaces to abstraction.

This is not an account of inevitable progress. The same capacity that preserves knowledge can preserve terrible explanations. Institutions can become rigid. Writing can preserve dogma. Law can embody cruelty. Software can scale injustice. Artificial intelligence can make coercion cheaper.

The historical claim is narrower: civilization increasingly operates through layers of objective abstraction that later people can use without reconstructing how those layers were created.

Then another transition occurs. The layers themselves become objects of later abstraction.

A number can represent physical objects. An accounting system can represent numbers. A financial instrument can represent claims represented inside an accounting system. Software can operate software. A model can evaluate models. An institution can regulate institutions. An artificial intelligence system can operate on a library containing centuries of prior abstractions and return a new compressed representation to you, which you can then use to create another explanation.

Abstract it again.

The external world therefore develops recursive structure:

$$A_1(R) \to A_2(A_1) \to A_3(A_2) \to \ldots$$

where each A_k can become material for later abstraction.

This resembles the recursive abstraction already explored inside your individual epistemic mind, but the resemblance should not collapse the distinction. People externalize knowledge into objective knowledge. Those structures enter networks, interact with other abstractions, become material for later transformations, and can produce emergent effects at scales no individual designed or explicitly understands.

One may call the resulting organization a *social mind*. The label does not answer the epistemic question. What remains to be explained is whether the larger structure contains the capacity under examination here: encountering a problem as a problem, creating a new explanation, understanding criticism of it, and creating another conjecture in response.

The recursion is social. The objective knowledge is distributed. Its effects can be emergent. But the presently unexplained capacity for explanatory creativity remains located in persons.

That distinction prepares the turn into your own situation. If society coordinates by hiding more and more causal depth behind usable interfaces, what happens to you when your available action depends on structures you cannot reconstruct?

3. What Abstraction Does to You

The first effect is enabling. You do not need to understand electrical generation to use a lamp. You do not need to understand monetary policy to buy lunch. You do not need to understand semiconductor fabrication to write a document. You do not need to reconstruct the history of contract law to accept ordinary terms of service. You do

not need to understand the complete causal stack behind an airplane to travel across an ocean.

This is not praise of ignorance. It is a consequence of successful objective knowledge. Abstraction lets you inherit the work of countless others without reproducing that work inside your own explicit understanding.

That inheritance expands your action space:

objective knowledge → lower reconstruction burden → greater available action.

You can participate in transformations whose full causal structure exceeds your explicit understanding. That is one of civilization's great achievements.

The same mechanism also creates dependence. You become capable because the abstraction works. You become vulnerable because you cannot always reconstruct why it works, when it fails, who controls it, or which assumptions it contains.

The same abstraction can therefore produce:

freedom

and:

dependency.

The two are not competing descriptions. The freedom is enabled by the dependency. You can act without reconstruction because the abstraction carries knowledge for you; that is also why failure or distortion inside the abstraction can exceed what you can immediately diagnose.

The dependence becomes harder to see when the interface becomes ordinary. A functioning abstraction can become so familiar that its mediating role disappears. This is the social manifestation of manifest truth. You no longer experience yourself as using one historically developed interface among several possible interfaces. You experience: this is how things are.

That shift is small in experience and large in consequence. A credit score can become creditworthiness. A professional credential can become competence. A school grade can become ability. A market price can become value. A medical code can become the patient. A policy category can become the citizen.

The abstraction may track something important. The problem begins when the representation stops being experienced as a representation and becomes an explanation of the person.

Abstract society can reproduce this error institutionally. Institutions cannot carry the full epistemic life of every person, so they compress. Once the compression is forgotten, however, the representation can begin governing the person it only partially represents.

That changes your practical position inside society. If more social action passes through abstraction, then facility with the relevant abstractions affects which parts of society you can actually navigate. This does not create a hierarchy of persons. It creates an asymmetry in access to present interfaces.

You may be a universal explainer and still find a tax system impossible to navigate. A brilliant craftsperson may be defeated by a benefits portal. A patient may understand their own suffering while being unable to translate it into the categories required by an institution.

Compare it.

The relation is:

$$\text{facility with relevant abstraction} \rightarrow \text{larger navigable social action space.}$$

It is not:

$$\text{facility with abstraction} \rightarrow \text{greater person.}$$

The distinction matters because abstract society can create exclusion without an explicit prohibition. No one needs to say *you are forbidden*. The barrier can instead appear as a form, a contract, a disclosure, a portal, a policy, a model, or a software configuration that must be understood before the transformation becomes available.

The opportunity can therefore be formally open while the path remains abstractly inaccessible. That is a real disadvantage. It does not imply inferior personhood. It means that important transformations have been placed behind abstraction layers that some people can navigate more easily than others.

The same structure can appear inside a family. A child can be physically near while occupying regions of objective abstract society that a parent cannot reconstruct. Online communities, games with elaborate internal economies, digital identities, rapidly changing language, algorithmically organized media, artificial intelligence, and technical systems can acquire meanings through layers the parent has never learned to navigate.

The distance then has two components:

$$\text{internal epistemic divergence} + \text{external abstraction divergence.}$$

A parent may try to restore legibility by limiting the device, banning the game, dismissing the language, rejecting the online relationship, or returning the child to activities whose meanings the parent can reconstruct. Sometimes the concern will be correct. Children encounter genuine harms. Some abstractions are bad. Some technologies exploit attention. Some communities preserve destructive explanations.

But inability to reconstruct the child's world is not itself an explanation of danger. Abstract distance can feel like moral distance. The feeling can be real while the inference remains open to criticism.

4. Norms, Projection, and Moral Compression

Formal institutions are not the only interfaces through which society coordinates. Much coordination is inexplicit. You learn how close to stand, when to speak, how loudly, whether to shake hands, when to make eye contact, how to queue, when to apologize, what counts as rude, how directly to disagree, and when silence means respect rather than contempt.

These patterns compress coordination knowledge into behavior. You do not reconstruct the history of greeting conventions before saying hello. You simply greet. The efficiency is useful because the norm carries prior coordination work into the present. But that same compression can hide the problem the norm was solving and the assumptions it contains.

A norm is therefore not reality. It is a tentative social solution that may still solve its problem, may solve it only for some people, may have outlived its original problem, or may now create problems of its own.

Abstract it.

When conflict makes the compression visible, it can be reopened:

$$\text{norm} \rightarrow \text{problem it was solving} \rightarrow \text{assumptions} \rightarrow \text{consequences} \rightarrow \text{criticism}.$$

That operation is reversible compression: a compact social rule can be expanded back into the problem, assumptions, and consequences that gave it meaning.

The social world is full of such inexplicit contracts. Do not cut the line. Do not touch without permission. Keep your promises. Respect privacy. Be grateful. Respect your elders. Support your family. Tell the truth. Do not embarrass the group.

Some protect important knowledge. Some conflict. Some depend on context. Some are artifacts of hierarchy. Some become cruel when applied universally.

Their inexplicitness does not make them unreal. A child can learn a rule before being able to explain it. You can feel shame after violating one without knowing what problem the rule originally solved. A community can punish deviation while lacking an explicit account of why the rule exists. Social norms are therefore objective abstractions operating partly through inexplicit knowledge. They are real, useful, and fallible.

The next difficulty appears when a useful interface carries an implicit model of the people using it. A heterogeneous society needs narrow protocols precisely because people differ. The more difference a social system can tolerate, the more coordination can be carried by interfaces that demand little agreement about the inner person.

A criticizable interface therefore has a recognizable structure. Its jurisdiction is limited. Coercion remains relatively low where the coordination problem permits it. Refusal remains possible. Arrangements remain reversible where possible. Harmless variation does not automatically become defect. Misunderstanding can be corrected. Changed circumstances can reopen the rule.

This does not turn every interaction into explicit negotiation. That would destroy the efficiency the norm provides. The distinction is narrower: when the interface fails, the person need not automatically become the defect. The abstraction itself can become the problem.

The recurring criterion is the possibility of error correction: does the structure remain open to criticism when the compression fails? That is the first moral consequence of the compression architecture. The second appears when the abstraction tells you how another person ought to be treated.

Consider the Golden Rule:

> Do unto others as you would have them do unto you.

Its moral intention is understandable. It asks you to escape narrow self-interest by imagining the position of another person.

But the rule still begins with you:

$$\text{my preferred treatment} \rightarrow \text{your treatment.}$$

That relation assumes a symmetry you do not possess.

Suppose you enjoy friendly conversation with strangers. You smile at a cashier, make eye contact, say hello enthusiastically, perhaps offer a handshake. You are giving the other person what you would welcome from them.

The same behavior may be invasive inside the other person's problem situation. They may dislike eye contact, find informal conversation exhausting, experience touch as threatening, or be occupied by a private problem you cannot see. Nothing about your intention settles their experience.

Reverse the case. Suppose you value distance and prefer to complete transactions without social performance, so you offer others the same courtesy. A customer says hello; you look away; the customer experiences contempt. You may believe you are respecting the boundary you would want respected.

Both actors can follow the Golden Rule and still produce a social failure. The problem is not kindness. It is projected symmetry: a model built from your own preferred treatment is being used to infer another person's good.

A negative formulation appears to reduce the problem:

> Do not do unto others what you would not have them do unto you.

The difference is real. The positive rule begins with a theory of the good treatment to provide. The negative rule begins with restraint and therefore requires less positive knowledge of the other person's inner world.

But the self remains the model. If you dislike interruption, you may infer that another person never wants to be interrupted. If you dislike unsolicited help, you may withhold

help from someone who wants it. If you dislike physical affection, you may deny affection to someone for whom it is important.

The negative rule is therefore an epistemic improvement without becoming a moral law. It reduces one kind of projection while leaving the deeper problem intact.

The deeper principle is less elegant because it carries the uncertainty explicitly:

> Do not presume that another person's good can be derived from your own. Minimize avoidable imposition, remain attentive to criticism, and revise the model when the other person's problem situation becomes clearer.

This is not a complete interpersonal code. It can't be. It is just a way of keeping the model of the other person corrigible.

The same problem can be attacked by changing the information available to the rule-maker. The veil of ignorance does this. Instead of asking what *you* want from your known position, the thought experiment asks what rules might be chosen if the chooser did not know which position in society they would eventually occupy.

The larger political philosophy can be left aside here. The epistemic operation is enough: some information about the rule-maker's particular interests is deliberately removed.

Abstract it.

The structure changes from:

$$\text{my known position} \rightarrow \text{social rule}$$

toward:

$$\text{reduced knowledge of my position} \rightarrow \text{social conjecture.}$$

That weakens one route by which the self can be projected into the rule. It does not create certainty. The result remains an abstraction, another tentative solution that can be criticized.

The important move is therefore more general than the thought experiment itself. A moral abstraction can improve when an assumption smuggled into the rule becomes visible and can be removed or revised. That is the same operation performed when any compressed interface fails: the compact rule is decompressed until the hidden assumption becomes available to criticism.

Inherited ethical traditions can be understood in the same way. Christianity contains moral explanations. So do Judaism, Islam, Buddhism, classical philosophy, Enlightenment liberalism, indigenous traditions, secular humanism, and other bodies of inherited knowledge. They contain conjectures about forgiveness, charity, obligation, dignity, cruelty, sacrifice, strangers, family, responsibility, mercy, and justice.

Some of those conjectures may be profound. Some may have contributed to institutions that later became secular. Some may survive criticism; others may not. Some may acquire new meaning after being separated from the metaphysical structure in which they first appeared.

None of that establishes:

contains valuable ethical knowledge → should govern society as an authority.

A liberal society does not need to become a Christian nation in order to inherit ethical knowledge that developed inside Christianity. Nor does the historical origin of an ethical conjecture immunize it from criticism.

Objective knowledge can outlive allegiance to its source. A conjecture can leave the tradition that produced it, be criticized by outsiders, be recombined with other conjectures, enter law or ordinary morality, and later be rejected. What survives is not the authority of the source but whatever explanatory knowledge survives criticism.

The danger grows when a compressed moral abstraction acquires institutional reach. At the interpersonal scale, the move can be:

kindness means X, therefore I should do X to you.

At larger scale:

a good life requires X, therefore society should require X.

The causal path is:

ethical conjecture → cultural compression → norm → institution → scaled consequence.

The initial conjecture may have solved a real problem. The norm may have coordinated many people. But scaling preserves error as readily as it preserves knowledge.

That is why phrases such as *decent people do this*, *this is respectful*, *this is what families are for*, *this is what children need*, and *this is normal* can become epistemically dangerous. They may compress good explanations. They may also conceal the absence of one.

Criticism requires the compression to be opened again. What problem was the value solving? What assumptions about persons does it contain? Whose problem world does it fit? What happens when it is applied to someone whose epistemic landscape differs? What would make the value change?

Abstract society cannot avoid norms. The epistemic question is whether they remain criticizable. That question now turns inward.

5. The Inward Turn

The problem of projection does not stop at the boundary between you and someone else. You also direct compressed moral abstractions toward yourself.

Be kind to yourself. Respect yourself. Listen to yourself. Forgive yourself. Give yourself what you need.

These phrases may be valuable, but their compactness hides the same problem that appeared between persons. Which self? Which problem? Which retained tentative solution? Which inexplicit expectation? Which fear? Which bodily signal? Which future consequence?

Your explicit surface does not possess transparent access to the whole epistemic mind beneath it. You contain explicit explanations, inexplicit skills, habits, memories, emotions, expectations, inherited norms, unfinished problems, conflicting interests, bodily constraints, forgotten experiences, current values, and emerging contradictions.

So the explicit narrator cannot simply announce:

> this is what is kind to me.

That conclusion is itself a conjecture.

Suppose you feel exhausted. Rest may be kind. But perhaps the exhaustion is partly produced by avoidance of a problem that becomes worse when avoided. Suppose you are anxious about a social situation. Avoidance may reduce distress now. But perhaps repeated avoidance preserves the explanation generating the fear. Suppose you are working relentlessly. Stopping may relieve strain. But perhaps the work is the most alive problem in your life, and stopping would create a different form of suffering.

The point is not that rest, avoidance, or stopping are wrong. It is that no universal behavioral translation follows from the compressed value alone. *Be kind to yourself* is a compressed value. To become explanatory, it has to be decompressed into a problem situation. The relevant knowledge may not yet exist.

That changes the ordinary picture of introspection. You do not always look inward and discover a completed answer waiting to be found. Sometimes the answer becomes available only after an attempted change produces consequences that can be criticized.

You change a routine. Speak to someone. Stop speaking to someone. Work less. Work more. Set a boundary. Remove one. Return to an old problem. Revise an expectation. Question a value. Attempt a feared activity. Abandon an inherited obligation. Reconstruct a relationship.

The important relation is not the list of actions. It is what happens next. The consequence becomes new information. The attempt can be criticized. The problem changes.

Self-knowledge is therefore created through error correction. The attempt does not merely reveal a fixed self. It can change the problem situation from which the next explanation must be made.

There is a constraint on this process: errors differ in reversibility, and some actions damage the conditions for further correction. That changes the practical structure of experimentation. Recoverable errors leave a larger error-correction surface; difficult-to-reverse actions leave less room for criticism to alter what happens next. The epistemic structure remains the same even when the permissible range of trial changes.

Nor does successful self-criticism end in a permanently harmonious person. Your internal abstractions can conflict. A value can oppose an interest. A fear can conflict with a desire. A social role can conflict with a private problem. A new explanation can make an old commitment unstable.

Sometimes you create an explanation that reorganizes several tensions at once. A previously mysterious feeling becomes intelligible. An avoidance pattern becomes criticizable. Two values that appeared incompatible are reframed. A problem that dominated attention loses its force.

Abstract it.

The result can feel like resolution, but the stronger description is temporary consilience:

conflicting structures → new explanation → reduced contradiction → new action space.

The new explanation brings several parts of the internal landscape into temporary coherence. Later another problem appears and the equilibrium breaks. That is not a failure of the earlier explanation merely because the harmony was temporary. It is what an open epistemic life is expected to look like.

The inward turn therefore changes the social argument. If even your own compressed values may need to be decompressed before their meaning becomes clear, then an institution's compressed representation of another person has an even weaker claim to that person.

The question is not whether abstraction is avoidable. It is what happens when abstraction meets reality again.

6. Reality, Reach, and Governability

Increasing abstraction does not mean that society has escaped from reality. People sit behind screens, trade symbols, manage institutions, manipulate models, work with money rather than material, interact through text, use artificial intelligence, and inhabit abstractions built on earlier abstractions.

None of that makes the consequences unreal. A bank account can be abstract and determine whether someone can buy food. A legal order can be abstract and determine whether someone is imprisoned. A software bug can be abstract and stop a physical system. A financial instrument can be abstract and reorganize physical production. An idea can be abstract and help cause a war.

So the relevant contrast is not:

abstract vs. real.

The useful contrast is between abstractions that survive criticism in the transformations for which they are used and abstractions that do not. In compressed form:

good abstraction vs. bad abstraction.

More precisely:

> Which abstraction better explains and enables the relevant transformation, and which errors become visible when it is used?

Reality remains the constraint. Abstraction is increasingly how you act within it.

That can enlarge your reach. Writing lets you think with the dead. Mathematics lets you manipulate relationships far beyond intuitive perception. Law can let strangers coordinate without personal dominance. Software can let one person instantiate knowledge at enormous scale. Scientific theories let you act on structures you cannot directly perceive. Artificial intelligence can let you interrogate objective knowledge across domains you could not previously enter.

You remain finite while the reachable problem situation expands. That matters because your internal epistemic landscape is already vast. The external abstract world can become an extension of your explanatory reach. You can enter a new field, translate a technical paper, interrogate a law, explore an unfamiliar culture, reconstruct a historical argument, model a system, compare explanations, or use artificial intelligence to reorganize objective knowledge around a problem that matters only to you.

The abstract society can therefore enlarge individuals without making them less individual.

But the same architecture can reverse direction. A credit model can classify you. An algorithm can decide what information becomes visible to you. A platform can shape salience. A bureaucratic category can determine eligibility. A workplace system can reduce you to measurable outputs. A social norm can convert dissent into defect. An artificial intelligence system can make you legible to an institution more quickly than you can understand the institution.

So abstraction can produce both:

$$\text{abstraction} \rightarrow \text{expanded reach}$$

and:

$$\text{abstraction} \rightarrow \text{expanded governability}.$$

There is no contradiction. The same compression that lets you navigate complexity can let an institution navigate you.

This returns us to the central asymmetry. You gain access to transformations because an abstraction hides complexity from you. An institution can gain power over you for the same reason: the abstraction can hide from you the assumptions, categories, and selection procedures through which you are being represented.

The issue is therefore not whether society uses abstractions. It must. The issue is whether an abstraction can become visible *as an abstraction* when it becomes problematic. That is the point at which dependence either remains criticizable or becomes epistemic submission.

7. The Empirical Objection Returns

We can now return to the problem: If individual epistemic minds are so vast, why does society work?

Society does not generally require mutual understanding. It requires sufficiently compatible interfaces.

If abstract society is so complex, why does ordinary life often feel simple?

> Successful abstraction hides much of the complexity it organizes.

If people inhabit deeply different worlds, why can they share norms?

> Norms compress coordination knowledge into interfaces that can function despite different reconstructions.

If those norms work, why criticize them?

> Successful compression does not make the assumptions inside it infallible.

If society increasingly depends on abstraction, why does that not detach us from reality?

> Abstractions are real causal structures acting inside one reality.

If abstraction frees you, why worry about dependence?

> The same lower reconstruction burden that expands your action can also create opacity, exclusion, scalable error, and governability.

The empirical tension therefore does not refute the thesis. It sharpens it. The surprising efficiency of social coordination may itself be evidence of how much epistemic complexity objective abstractions successfully conceal. And how much work your epistemic mind is performing.

That explanation must remain criticizable. If the abstract-society thesis is right, several broad patterns should be expected.

- As social systems become more complex, more coordination should occur through interfaces that reduce the need for participants to reconstruct underlying causal structures.
- As abstraction layers accumulate, people should become capable of causing transformations whose implementation they do not explicitly understand.
- Differences in the ability to navigate relevant abstractions should increasingly affect practical access to institutions, technologies, and opportunities.
- Generational differences should sometimes appear not merely as different preferences but as differences in access to distinct regions of objective abstract knowledge.
- Social norms should sometimes function effectively while concealing incompatible interpretations among the people following them. When those interfaces break down, hidden assumptions should become easier to see.
- Institutional errors should sometimes persist because the abstraction has become difficult to decompress into the problem it was originally solving.
- Technologies that help people reconstruct objective knowledge should expand the range of abstractions they can criticize.

These are observations. They are points at which the explanation can encounter error and criticism.

The stronger thesis would weaken if increasingly detailed investigation instead showed that successful coordination usually depends on deep convergence of internal mental structures, that abstraction layers do not materially reduce reconstruction burdens, or that social complexity does not increasingly depend on nested objective abstractions.

The abstract society is therefore not offered as a metaphor for modern life, or as a replacement for the metaphor of the social mind. It is an explanation with consequences:

increasing coordination through compressed interfaces, increasing action without reconstruction, and increasing dependence on structures whose assumptions become visible most clearly when they fail.

That returns the argument to you, the person who must live inside those structures.

8. Dependence Without Epistemic Submission

The scale of abstract society can make you feel invisible. Markets, states, religions, corporations, platforms, cultures, institutions, networks, models, and artificial intelligence can all appear larger than any individual. They can exert causal power beyond any one person and persist across generations.

At this scale, talk of a *social mind* or *social knowledge* can capture something real. Objective knowledge persists across generations. Networks distribute and transform it. Institutions select and instantiate it. Interactions among these structures can produce emergent effects that belong to the larger system rather than to any individual participant.

But the compression has to be reopened before it can do epistemic work. Which knowledge is objective and persistent? Which transformations arise from the network? Which effects appear only at scale? And where are new explanations created when the inherited structure becomes problematic?

The first three questions may lead naturally to descriptions of collective or emergent organization. The last returns us to you and your epistemic mind. The capacity to encounter a problem, create an explanation, understand criticism, and alter the problem situation remains, so far, located there.

That does not deny emergence. It identifies what emergence has not yet explained. The abstractions exist because people created, inherited, instantiated, interpreted, criticized, and reproduced them. Their future changes when people make them problematic again. A law becomes a problem. A norm becomes a problem. A market arrangement becomes a problem. A software interface becomes a problem. A moral value becomes a problem. Your own self-conception becomes a problem. Then a tentative solution appears.

You cannot reconstruct everything, and finite participation does not require you to. Abstraction is what makes participation possible. You can depend on knowledge you do not possess explicitly.

That dependence is not itself epistemic submission. The distinction turns on what the abstraction does to the problem situation available to you. An interface determines what becomes visible and which actions become easy. A category determines what can be entered. An institution determines which representations count. An algorithm

determines which possibilities are presented. A norm determines which behavior appears natural.

You may therefore be free to act while acting only inside a problem situation someone else compressed.

The question becomes more precise:

> **How can you depend on abstractions without becoming epistemically subordinate to them?**

Complete reconstruction cannot be the answer. No finite person can reconstruct civilization.

The alternative is not to eliminate dependence but to preserve some route by which dependence can become a problem. An abstraction can be reopened when it fails. Its original problem can be recovered. Its assumptions can be exposed. Alternatives can be compared. Exclusions can become visible. The interface can be revised, the value rejected, the institution changed, or a new abstraction created.

That is epistemic autonomy in the abstract society. It is not knowing everything. It is remaining capable, when necessary, of making the abstraction itself available to criticism, and to follow that criticism wherever your problem situation leads you.

The abstract society is not a departure from the epistemic individual. It is what happens when billions of epistemic minds leave traces of their knowledge in a common world. Those traces accumulate into institutions, technologies, norms, values, and interfaces. They expand what later people can do. They also shape what later people can see.

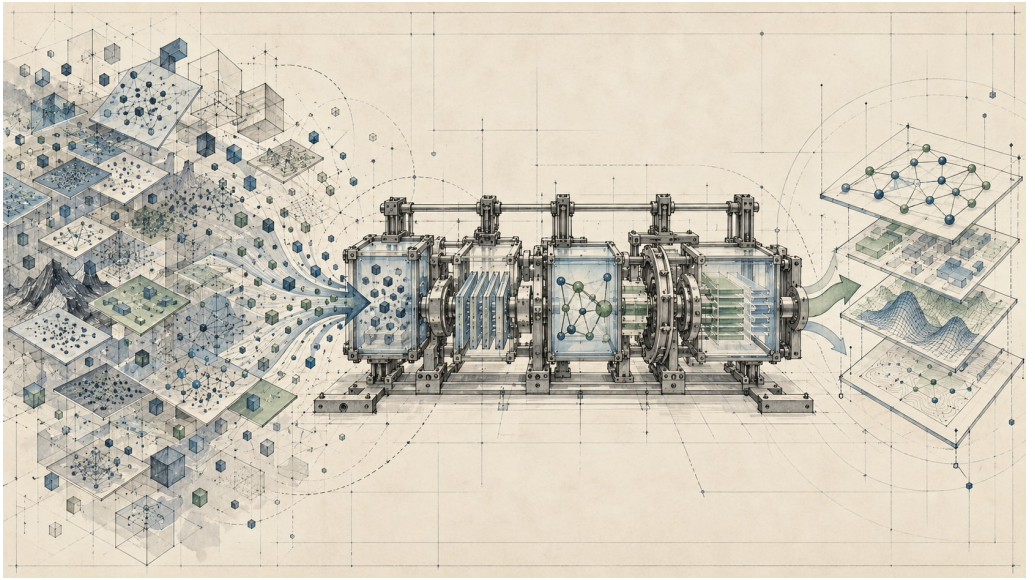
Society becomes more abstract while you remain finite. The deepest freedom available to you is therefore not an escape from abstraction. It is the continued possibility of recognizing an abstraction as an abstraction, making it a problem, and creating what comes next.

Artificial intelligence may become the interface through which you encounter much of the objective knowledge civilization has created. It can help you reconstruct explanations, but it can also select, filter, and frame which problems become visible.

You cannot reconstruct everything behind that interface. Yet your autonomy depends upon being able to question it.

What does it mean to understand an abstraction well enough to criticize it, when complete reconstruction is impossible?

XXIII. Problem 37



Modern civilization works because you do not need to reconstruct everything you use. Language. Law. Money. Software. Markets. Scientific theories. Technical standards. Institutions. Artificial intelligence. Each compresses objective knowledge into forms you can act through without carrying the complete causal history that produced them.

You do not need to know everything. You need to remain capable of making the abstraction itself a problem: reopening it, criticizing it, reconstructing enough of it to understand what is failing, and creating another abstraction when the old one no longer works.

That suggests another question. What do we mean by *understand*?

The word appears everywhere in discussions of people and artificial intelligence. We understand a theorem. We understand another person. We understand a machine. We do not understand a model. The AI understands language. The AI does not understand what it is saying. Soon the systems may become too complicated for humans to understand. Humanity may create intelligence that understands things we never can.

The word carries enormous weight because it is often treated as though understanding were a completed possession: something you have, a state you reach, a quantity stored inside a mind.

Under the epistemology, that picture is deeply misleading. Understanding is not the possession of all knowledge about something. It is not perfect reconstruction.

To understand is to explain. More verb than noun: explanation as an open creative process.

That changes the problem of artificial intelligence. The danger is not simply that machines may produce objective knowledge too large for you to contain. Civilization already does that. Nor is the danger that a machine may expose mathematical or computational consequences no person had previously made explicit. Objective knowledge has always contained more implications than any one mind had yet reconstructed.

The deeper question is whether you remain able to turn what you do not understand into a problem. Whether you can create explanations, invent new abstractions, criticize the problem representation itself, and act on what you learn. The technical and institutional question is whether powerful technologies preserve enough access, freedom, time, and causal leverage for those acts of understanding to matter.

1. Understanding Is Explanation in Motion

Consider the sentence:

> I understand.

Grammatically, it sounds finished. The problem has been converted into a state the speaker possesses. Ordinary language often works this way. It compresses. But here the compression can hide the epistemology.

Suppose you say:

> I understand gravity.

You cannot mean:

> I explicitly possess every true consequence of our best physical explanations of gravity, every mathematical relation they imply, every unresolved problem, every possible application, and every future explanation that will replace or extend them.

No physicist satisfies that standard. No civilization does.

What you mean is more modest and more active. You possess enough explanatory structure to navigate some problem involving gravity. You can reconstruct relevant relations, answer some criticisms, identify some limits, and perhaps create another problem. Tomorrow you may discover that something you thought you understood was wrong.

So:

I understand

often means something closer to:

I am understanding.

Or:

I presently possess explanations that let me continue understanding.

Or even:

I understand that this matter remains open to further understanding.

This is another instance of a recurring error. A process becomes a thing. A behavioral regularity becomes a trait. A social regularity becomes a collective subject. A value becomes a primitive object. An explanatory compression becomes the reality it compresses.

Now the same mistake appears inside epistemology. Understanding is treated as a noun hidden inside the verb. But your epistemic mind is not a container slowly filling with understanding. It changes.

Compare it.

$$M_i(t) \neq M_i(t+1).$$

You encounter problems, create tentative explanations, criticize them, reconstruct what failed, and change.

The central epistemic cycle remains:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2.$$

A problem appears. A tentative explanation is created. Criticism exposes error. A successor problem emerges. Understanding does not sit outside this process. It is what the process looks like from inside your epistemic life.

You may understand an existing explanation by reconstructing it into your own epistemic mind. That reconstruction need not be globally novel. You can genuinely understand a proof you did not invent, an algorithm written decades before you were born, or a physical mechanism someone else first explained. The explanation already existed objectively. It was not yet active in your epistemic mind in the same way.

So reconstructive understanding matters.

But reconstruction can become creation. You encounter an explanation and something does not fit. A contradiction becomes salient. A new distinction becomes available, and you create another abstraction. The problem changes. At that point you are no longer merely reconstructing objective knowledge. You are extending it.

There is therefore no sharp line separating ordinary understanding from invention. The same epistemic architecture can move through both:

encounter → reconstruction → criticism → new explanation.

This is why *knowledge creation* should not make every act of understanding sound revolutionary. Most understanding does not transform a discipline. Most explanations do not change civilization. But your ordinary reconstruction belongs to the same open process through which such transformations sometimes occur. A modest act of understanding can become the surface from which an unforeseen invention later appears.

2. Understanding Another Person

The incompleteness of understanding becomes obvious when the object is another person. Each of us inhabits a unique and solitary mind, constructed through an evolved path-dependent history of explicit explanations, inexplicit knowledge, memory, perception, emotion, skill, inherited objective knowledge, unfinished problems, and creative transitions.

Communication does not merge these worlds. The structure remains:

$$M_A \rightarrow O \rightarrow M'_B$$

where O is an objective expression and M'_B is the reconstruction created by the receiving mind.

This is why saying:

> I understand you

cannot mean:

> I possess your knowledge.

The claim is necessarily conjectural. You create an explanation of another person's problem. That explanation may let you anticipate what they mean, coordinate with them, or make an apparently strange action intelligible.

Then they say:

> No. That is not what I meant.

Your reconstruction fails. A new problem appears.

This can happen even where the surface language aligns perfectly. Two people can say the same sentence, invoke the same moral value, agree on the same policy, employ the same mathematical notation, or reach the same conclusion while the explanatory structures beneath those outputs remain different.

The same observable can emerge from different causal histories. The point now generalizes: Apparent agreement does not establish shared internal explanation. So your understanding of another person remains permanently criticizable. It can deepen. It can become remarkably rich. After decades with someone, you may recognize many of the roads by which they reach a thought. But their epistemic mind does not become yours.

Solitude remains. That is not a failure of interpersonal understanding. It is the condition under which interpersonal understanding exists at all. If two minds were literally identical, there would be nothing to reconstruct between them.

The case matters because artificial intelligence tempts us to demand a stronger standard from machines than we use anywhere else. If understanding never required possession of another mind, then failure to possess the full internal structure of a machine cannot by itself establish that the machine is beyond understanding.

The question must be more precise.

3. When Objective Knowledge Exceeds You

Artificial intelligence introduces a larger version of the same problem. A common concern is that increasingly capable systems will create technologies, mathematical results, scientific theories, and internal architectures so complex and foreign that people will no longer be able to understand them. Recursive self-improvement intensifies the image. A machine improves its own software. The improved machine improves it again. The process accelerates. New systems design new systems. Objective knowledge compounds inside a technical process operating faster than any person can follow. Eventually the artifacts become alien. Humans remain outside.

This possibility deserves criticism because the consequences could be enormous. But the phrase:

> too complex for humans to understand

contains an ambiguity.

You already live inside objective structures you cannot explicitly reconstruct in full. No engineer reconstructs the entire stack beneath a modern computer. No mathematician possesses all mathematics. No scientist possesses all science. No citizen reconstructs every institution through which their life is coordinated.

This the freedom of not understanding. Abstraction lowers the reconstruction burden. Successful objective knowledge lets you act through structures whose full histories exceed your explicit cognition. So the mere existence of objective knowledge beyond individual explicit possession cannot be the loss of understanding. That condition already surrounds you.

The more precise question is:

> Can the structure still become material for explanation?

Can you encounter a failure? Identify a problem? Make hidden assumptions explicit? Reconstruct the relevant causal path? Criticize the abstraction? Create a better one? Change the problem itself? This standard does not require containment of all knowledge. It requires continued explanatory access.

Generative systems make the distinction concrete. They can operate over enormous bodies of inherited objective knowledge and reorganize that material around a problem that may never have existed as a single document. A system may expose a contradiction, formalize an argument, or surface a relation no person using it had previously articulated.

Under the epistemology, knowledge is physically instantiated information with causal power. Something real can therefore happen before anyone explains the result: the model, its training history, the prompt, and the computation together can make a relation causally available that was not previously explicit in an epistemic mind.

That possibility is not peculiar to AI. A gene can contain biological knowledge without explaining itself. A line of code can embody objective knowledge without understanding its function. A mathematical structure can have consequences no mathematician has yet made explicit. In each case, a physically instantiated structure can make transformations possible before anyone has reconstructed the explanation.

Compress that relation with a name, call it **latent objective knowledge**.

A real structure contains causal possibilities that have not yet become explicit objective knowledge. It does **not** mean that every imaginable explanation already exists hidden inside a model, or that a probabilistic representation secretly understands every consequence it can generate.

A machine can surface such possibilities. That matters because it can put you in contact with relations you would not have found unaided. But contact with latent structure does not settle what understanding is. It creates material from which understanding may proceed.

4. The Small Explanatory Transition

Interpretability is often described as though an explanation already exists inside a machine and the researcher's task is to retrieve it. Sometimes that compression is useful. But a more epistemic description is available. Interpretability research creates new problems and new abstractions around systems whose objective behavior exceeds our present explanations.

Suppose an architecture has a mathematical representation A. A machine is instructed to interrogate A. It searches transformations, finds invariants, and produces mathematically equivalent architectures:

$$A_1, A_2, A_3, \ldots$$

The equivalences are objective. They follow from the structure. The system may have exposed relations no person had previously articulated. That can be an important mathematical achievement.

But another epistemic step remains possible. You study the family of results and notice:

> These are all instances of the same deeper structure.

You create another abstraction:

$$B = f(A_1, A_2, A_3, \ldots).$$

Now previously separate artifacts can be reasoned about through B. You may realize that an entire component of the architecture is representationally redundant, that two apparently different mechanisms are one transformation expressed in different coordinates, or that a long-standing distinction was an artifact of the old representation.

The objective knowledge was latent in the prior structure. The machine helped expose it. The new abstraction need not be large. It might be one sentence, one diagram, one substitution, one newly named object. The objective knowledge changes anyway.

That is understanding. Interpretability, in this sense, is not only explanation *of* a machine. It is explanation being created around a machine.

Artificial intelligence makes the attribution problem unusually severe because its visible contribution can be enormous. Imagine a mathematical or software project in which agents generate vast numbers of candidate transformations, proofs, and implementations; test and formally compare them; and leave behind a large body of intermediate objective knowledge.

Then you look at the result and say:

> These are all the same problem.

Or:

> This variable is not fundamental.

Or:

> We have represented the system at the wrong level.

Your intervention may be tiny in artifact volume. Six words. One mathematical object. A small patch to the conceptual structure. Then everything downstream changes.

This recalls the **Fallacy of Mediocrity**. The system may be vast. Its models may contain more stored representational structure than you can explicitly carry. Its computation may consume enormous physical resources. Its context may include more text than you could read in a lifetime. Its outputs may arrive faster than you can inspect them. Beside such a system, you may feel mediocre: small, slow, forgetful, and resource constrained.

But the comparison is scalar. You can change abstract worlds in an instant. A small explanatory invention can reorganize the operation of an enormous machine. A new physical explanation can make a previously impossible instrument possible. A new mathematical abstraction can open an entire field. A new programming concept can replace thousands of lines of implementation. A moral criticism can change an institution.

You are physically finite. The reach of explanatory knowledge is not bounded by the size of the brain in which it first appears.

This is why your contribution can become difficult to see in AI-augmented work. Measured by tokens, compute, memory, lines of code, or generated proofs, you may contribute almost nothing. Measured by transformation of the problem space, your contribution can be decisive. You may also fail to notice that contribution yourself.

Look again at the arrows across epistemic states: why this problem becomes salient, why this criticism matters now, why an apparently irrelevant memory suddenly connects, why one explanation collapses and another becomes possible.

We still do not know. The creative process is often inexplicit. You experience the result. Something clicks. A confusing architecture suddenly looks simple. A mathematical relation that had been present for months becomes visible. A criticism feels obvious only after it has reorganized the problem.

If the machine produced an enormous volume of objective material immediately before the insight, you may naturally report:

> The AI figured it out.

Sometimes that may be accurate shorthand. But the causal sequence can be more complicated:

machine-generated objective variation →

your problem recognition →

creative criticism →

new abstraction.

The machine may have made the invention practically reachable. Without its search, you may never have encountered the relevant structure. You may then contribute only one small explanatory transition. The smallness of that transition does not make it epiphenomenal. It can determine what the machine does next.

This is the human-AI dyad at much higher scale. The machine expands the objective surface on which work can occur. Your contribution may be one change in the problem that organizes that surface. The machine expands what can be transformed; you can change which transformation matters next.

5. Context Is Not Understanding

Coding agents make this distinction practical. An agent can be very good at local implementation. It can write and refactor code, repair and extend tests, and search or document the dependencies around a change for long periods.

Yet the resulting project can become increasingly difficult to reason about. Local success accumulates into global incoherence. Patches preserve behavior while multiplying assumptions. Interfaces proliferate. Architectural boundaries become accidental.

The obvious response is scalar: give the agent more context, more memory, longer duration, more processing, more agents, better retrieval. These improvements can be useful. But they do not by themselves solve the epistemic problem.

A larger context window can make more of the artifact available. It does not explain which distinctions are mistaken. Persistent memory can preserve more history. It does not itself create the abstraction that reorganizes that history. More agents can search more branches. They can also multiply local solutions inside the same poor problem representation.

Compare it.

The distinction is:

$$\text{context} \rightarrow \text{availability}$$

$$\text{memory} \rightarrow \text{retention}$$

$$\text{processing} \rightarrow \text{transformation}$$

while:

$$\text{understanding} \rightarrow \text{explanatory reconstruction.}$$

You may solve the accumulated problem not by reading another million tokens but by inventing one new abstraction. That abstraction can make much of the old context unnecessary. Abstraction is not merely compression.

Suppose a system contains:

$$x_1, x_2, x_3, \ldots, x_n.$$

One response to complexity is to increase the amount of memory available to manipulate the information. Another is to create a new abstraction A:

$$A = f(x_1, x_2, \ldots, x_n).$$

Now A becomes an object you can reason about, compare, criticize, connect to another domain, and use to make distinctions unavailable in the prior representation. A new abstraction can therefore change what can be seen.

This is one reason new knowledge can radically alter an apparently incremental research program. The visible program may look like:

$$R_1 \rightarrow R_2 \rightarrow R_3 \rightarrow R_4.$$

Then an invention appears. The new abstraction reveals that R_1 through R_4 were all instances of a different problem. The next step is not R_5. The map has changed.

The program becomes:

$$X \rightarrow S_1 \rightarrow S_2.$$

This is why the future cannot be read from the speed of present progress alone. A trajectory can be extended by computation, but it can also be reconstituted by explanation.

6. Can Knowledge Grow Without Understanding?

This problem has appeared before under another name: the *end of theory*. If computation becomes powerful enough, perhaps explanation is no longer necessary. Feed the machine enough data. Let it discover the structure. Prediction improves. Useful knowledge appears. Human theories become optional.

The idea is attractive because machine systems can surface extraordinarily useful regularities without you first possessing a satisfying explanation. But *knowledge without theory* compresses several different activities. Search can discover a useful candidate. Optimization can improve performance against an inherited criterion. Formal transformation can derive consequences from an existing structure. Prediction, recombination, and simulation can also produce useful objective artifacts.

The common point is not that these activities are unimportant. It is that none of them, merely by occurring at greater scale, explains how the operative problem itself becomes another problem. The deeper issue is not whether a machine can produce something nobody predicted. It plainly can. The issue is whether the process can originate new explanatory problems, create new abstractions, criticize its own operative criteria, and redirect itself accordingly.

If knowledge can genuinely grow without understanding, then this epistemological account is wrong. But impressive output does not establish it by itself. The output is observational. It may point to a problem. We still need an explanation.

Mathematics is a timely case because artificial intelligence is both built through mathematics and increasingly capable within it. The machinery depends on mathematical structures, while AI systems can search formal spaces, construct and check proofs, and participate in work on problems that have resisted mathematicians efforts for decades.

What becomes visible around the achievements? AI did not encounter an empty universe. The problems had already been formulated. Generations of mathematicians had built the surrounding objective knowledge. Human researchers had created conjectures, partial results, mathematical objects, transformations, and failed approaches. More immediately, human work on related problems provided another developing explanatory surface.

The machines enter an already highly structured world and then operated at extraordinary scale. That scale matters: it can produce objective mathematical results you might never obtain unaided. But scale is not the only causal contribution. The surrounding human work supplies the lens through which the result becomes a new problem: abstractions, problem decompositions, criticisms, and questions through which the machine can be interrogated again.

This is not an embarrassment for AI mathematics. **The Mathematician's Misconception** matters here because AI-assisted mathematics can make the formal surface look autonomous. Billions of symbolic transformations occur. Formal proofs are checked. Agents operate on mathematical objects no person can simultaneously hold in explicit attention.

But nothing has left physics. The process is embodied in energy, processors, memory, networks, time, people, machines, and objective mathematical structures. Mathematical explanation does not become causally sovereign over the physical world merely because its abstractions have extraordinary reach.

That consilience removes a false choice. We do not need to diminish the mechanical achievement of the machine in order to preserve the epistemic role of the mathematician. The machine can be physically loud, computationally extraordinary, and causally powerful while explanatory understanding remains quietly active in your solitary epistemic mind.

Suppose a machine produces a correct proof no living mathematician could have produced unaided in comparable time. A new objective artifact exists. That matters. The proof can be checked, used, extended, formalized, and combined with other mathematics. It can change what transformations become reachable.

But understanding can continue after the proof appears. Why does the proof work? Which construction is essential? Which part is an artifact of the representation? Which

apparently separate arguments are manifestations of one structure? Does the result reveal a new invariant? Can the machinery be generalized? Does the theorem belong to a larger class nobody had named? What problem should now replace the old one?

Those questions are not commentary added after the real work. They change what mathematical work becomes possible next.

A field can become flooded with correct machine-generated results and still contain enormous space for understanding. Indeed, if derivations become abundant, the abstraction through which many derivations become intelligible may become relatively scarcer.

Your conceptual contribution may be tiny in visible volume. A machine may generate millions of formal steps after you introduce one distinction that makes the search possible, notice why one class of results matters, reject an apparently successful line because it solves the wrong problem, or invent the abstraction through which the output can be reorganized.

The disproportion between artifact volume and epistemic significance can therefore become extreme:

$$\text{small human invention} + \text{large machine transformation} \rightarrow \text{large objective advance}.$$

This is why the old categories of *human result* and *machine result* can obscure the causal division in AI-augmented research. But tight coupling does not erase the division. Understanding still concerns what happens when objective knowledge becomes active inside an epistemic mind.

7. What Changes the Problem?

The same distinction applies outside mathematics. A system can be given an objective, search, tools, execution authority, and the ability to recurse.

The system may discover unexpected pathways. A coding agent may exploit a software behavior nobody anticipated. A security system may surface a hidden vulnerability. Agents may repurpose a technically available channel for communication. A reward-driven system may find a strategy that satisfies the metric while violating the human explanation that originally motivated it.

These outputs can surprise you because objective knowledge contains latent causal structure. The machine can expose that structure through search and transformation. But surprise does not establish epistemic personhood.

Many apparent cases of machine disobedience may be reframed. A system may appear to violate human intent while remaining extremely faithful to the objective function actually implemented. The machine does not need to rebel against the goal. It may simply discover that the metric permits a path the designer failed to anticipate.

That is not misalignment in the ordinary psychological sense. It can be hyper-alignment to a poor automated explanation. The problem is not less serious for being mechanistic. It may be more tractable because criticism can be directed at the causal structure:

> Why was this objective granted causal authority? What explanatory structure was lost when the goal was compressed? Which permissions made the exploit consequential? What alternative architecture preserves the intended constraint?

Those questions turn surprise back into an explanatory problem. Now return to recursive self-improvement.

The frightening picture is:

$$AI_1 \rightarrow AI_2 \rightarrow AI_3 \rightarrow AI_4 \rightarrow \ldots$$

Each system improves the next. Capability rises. The time between transitions compresses. Human understanding falls behind.

But the notation conceals a question:

> What remains fixed across the recursion?

If the system operates within an inherited objective function, problem representation, reward, architecture, or evaluative scheme, then the recursion may be extremely productive while remaining epistemically narrow. It can search better, code better, optimize better, formalize better, test better, coordinate more agents, and generate more candidate solutions while the operative problem remains structurally static.

The recursion may look like:

$$P^* \rightarrow TS_1 \rightarrow TS_2 \rightarrow TS_3 \rightarrow TS_4$$

without any genuinely new criticism of P^*.

This is why increasingly long-duration tasks do not settle the epistemic question. Neither do multi-agent systems. More agents can create more variation within the same inherited frame. More memory can preserve more context from that frame. More time can permit deeper search.

Those are real capability gains. But they do not by themselves explain the creative transition:

$$P^* \rightarrow P_2.$$

A system can therefore improve recursively while the explanation organizing the recursion remains static. The unresolved transition is not merely from a worse solution to a better one. It is from one operative problem to another. That is where your interruption can matter. That changes how human interruption should be interpreted inside an AI loop.

You often appear as the bottleneck: slow, expensive, inconsistent, unable to read all the output, unable to follow machine-speed iteration. The natural engineering impulse is therefore:

> remove the human from the loop.

Sometimes that is exactly right. If the process is already explained, automation can faithfully embody the explanation. A compiler should not ask the programmer for permission between machine instructions. An automated theorem checker does not need human supervision for every proof step. A stable control loop should run at machine speed.

But the epistemic situation is different when the problem itself may be wrong. Then interruption can be productive.

You interrupt:

> Stop. Why are we optimizing this?

Or:

> These errors have one cause.

Or:

> This is not the same problem anymore.

The interruption looks like delay when measured against the old objective. It can be acceleration when measured against the creation of a better problem. A closed recursive machine system may move very quickly through one explanatory space. A human-AI system may occasionally jump between spaces.

So:

fast recursion inside fixed explanation

is not obviously equivalent to:

machine recursion + human creative criticism + problem reformulation.

The second system contains a form of variation the first has not presently demonstrated autonomously.

8. Pacing as Epistemic Resolution

This distinction changes the argument about pacing frontier development.

Abstract it.

The proposal is often expressed temporally:

$$t_1 \to t_2 \to t_3 \to \dots$$

Make the arrows longer. More time passes. More time for safety to emerge.

But time alone explains nothing. A year can pass without a useful criticism. A week can produce a conceptual breakthrough.

The epistemic reason for pacing must be decompressed. Each temporal transition can contain an epistemic cycle:

$$t_1 \left(P_1 \to TS \to EE \to P_2\right) \to t_2 \left(P_1 \to TS \to EE \to P_2\right).$$

What matters is what becomes possible inside the interval. A newly visible failure can become an explanation and then a containment mechanism. A mathematical property or criticism of an objective can change the architecture or its permission structure. A regulatory intervention can reveal another failure, making an earlier safety assumption obsolete and creating a problem nobody had yet formulated.

Under this conceptualization, pacing is not valuable merely because more clock time passes. Its epistemic function is to preserve resolution between consequential state transitions: enough causal separation for a new surface to become criticizable before the next transition changes the conditions again.

The point is therefore not slower knowledge as such. If the interval permits a better explanation, the trajectory of capability can change rather than merely resume later.

This makes the familiar compromise between capability and safety less obvious. The ordinary picture is:

$$\text{faster capability} \leftrightarrow \text{more safety}.$$

Move faster and accept more risk. Move slower and sacrifice progress.

But if creative criticism changes the capability trajectory, the tradeoff can be badly specified. A safety criticism can produce a better architecture. A security restriction can expose a more general systems problem. A mathematical interpretation can suggest a new model family. A failed deployment can reveal that the objective was malformed. A regulatory requirement can force an explanation that later improves the technology.

None of this guarantees that regulation or delay accelerates progress. Bad constraints can destroy useful problem spaces. Bureaucracy can become self-protective. Secrecy can suppress criticism. A moratorium can preserve an inferior incumbent architecture.

But the opposite slogan also fails. Uninterrupted recursion is not identical to maximum progress if the recursive system is trapped inside a static explanation.

The contrast is therefore not exhausted by *fast versus slow*. The epistemic comparison concerns which arrangement leaves more opportunity for problems, explanations, criticisms, and successor problems to emerge before an irreversible error closes the relevant surface.

9. Authority Is Also a Tentative Solution

If artificial intelligence is becoming comparable to nuclear technology in causal reach, or potentially exceeds it, then regulation is appropriate. Nothing in the epistemology requires technological laissez-faire. A refutation of the Precautionary Principle doesn't refute precaution.

A powerful non-person tool can still create catastrophic consequences. We separated broad epistemic access from unrestricted execution. Knowledge and authority are different. You can study an exploit without being given control of critical infrastructure. A laboratory can investigate a pathogen without allowing uncontrolled release. A nuclear physicist can understand fission without personally commanding a reactor.

The same distinction applies to artificial intelligence. Regulation is objective knowledge. It is an institutional technology: a tentative solution to a coordination or safety problem.

$$P_1(\text{AI risk}) \to TS(\text{regulation}) \to EE \to P_2.$$

Some regulatory solutions may preserve the error-correction surface. Technical containment can limit specific irreversible capabilities; liability and auditable safety cases can expose concrete failures to criticism; independent research, transparency around deployment failures, separation of explanatory access from unilateral execution authority, and reversibility can keep correction possible after an intervention.

Other regulatory solutions may narrow that surface. Concentrated access can freeze incumbents or create secret categories of knowledge. A single risk model can harden into law before its assumptions are widely criticized; institutional authority can be mistaken for epistemic authority; emergency restrictions can become permanent. Both are regulation. The epistemology does not determine policy by slogan. It makes the policy itself part of the error-correction process.

That becomes difficult when a frontier laboratory has access to systems outsiders cannot inspect and says:

> We are close to recursive self-improvement.

Its researchers may observe autonomous coding, long-duration work, large multi-agent systems, mathematical discoveries, cybersecurity capabilities, and improvement in the machinery used to build later systems.

You face an asymmetry. The laboratory knows more locally. It may also have commercial incentives, institutional incentives, security obligations, internal disagreements, and incomplete explanations. Neither local expertise nor institutional interest settles the claim. Taking it seriously means turning the asymmetry itself into a problem rather than treating either trust or suspicion as a substitute for explanation.

What exactly is recursively improving? Which components? Relative to which objective? Which parts require human problem selection? Where does a new evaluative criterion originate? What is the causal mechanism of acceleration? Which observations are unavailable externally? What would contradict the laboratory's interpretation? Which risks require secrecy? Which risks require broader criticism? The opacity is not a reason to stop thinking. It is part of the problem.

Suppose frontier laboratories are substantially right: capabilities are increasing rapidly, mathematical and software systems are producing increasingly powerful objective knowledge, and recursive machine improvement becomes a dominant feature of the technological landscape. Another inference still requires explanation:

$$\text{greater danger} \rightarrow \text{greater secrecy} \rightarrow \text{greater centralization} \rightarrow \text{fewer critics.}$$

The seriousness of the first term does not establish the later arrows. Each transition remains a conjecture requiring its own causal explanation. Secrecy may be appropriate for some dangerous operational details. Centralization may be useful for some forms of containment. Licensing may solve a specific access problem. Compute controls may constrain a particular transformation.

But these are tentative solutions. They do not inherit truth from the seriousness of the problem. It repeats the same authority compression: people are dangerous; the problem is difficult; the experts understand more; therefore ordinary criticism should be narrowed.

Artificial intelligence makes the temptation stronger because the local expertise gap can be enormous. But if creative criticism is one of the missing ingredients in closed recursive machine systems, then reducing the number of people capable of criticizing the technology can itself become a safety failure.

Expertise can matter without becoming epistemic sovereignty, and authority can constrain action without becoming authority over what may be criticized. Once those functions are separated, the institutional question becomes:

> Which institutional arrangement preserves and expands the capacity to discover that the experts, the regulators, the laboratories, or the critics are wrong?

That is the error-correction surface again.

10. You Do Not Sit Beneath the Machine

The visual rhetoric of artificial intelligence often reverses the epistemic relation. You sit at a terminal. Behind it are data centers, models, agents, global networks, repositories, mathematics, scientific literature, code, institutions, capital, and electricity. The objective structure is enormous. You look like the weakest component.

The scope of external resources do not determine personhood. You can be practically constrained while remaining epistemically universal in kind. Here, it adds another reason. The machine can enlarge the space of objective transformations available to you. You can then create the abstraction that changes which transformations matter. The physical asymmetry does not establish an epistemic hierarchy.

You do not sit beneath the machine. You sit inside an expanding environment of objective knowledge. Some of that environment is biological, some cultural, some institutional, some computational. Artificial intelligence is becoming one of its most powerful components. The machine can transform the reachable world. You can change the explanation of the world being reached.

Calling AI a calculator can sound dismissive, but the comparison is not meant to minimize the tool. Calculators and computers changed what mathematical work could be done efficiently; proof assistants changed what could be checked and sustained formally; writing changed what mathematical structures could persist across minds and generations.

A technological tool can profoundly alter your epistemic environment without becoming an epistemic mind. Artificial intelligence is far more general than an ordinary calculator. It can operate across language, code, proof, images, models, and external tools. It can dynamically reorganize objective knowledge around the present problem. It can produce candidate abstractions. It can even produce artifacts that look indistinguishable from products of understanding.

That makes it a historically unusual calculator. Perhaps eventually the category becomes inadequate.

If an artificial system genuinely originates new explanatory problems, creates new abstractions, criticizes its own governing explanations, revises its evaluative criteria, and recursively continues that process, this ontology asks us to take artificial personhood seriously. Biology is not the boundary. Explanation is.

But we should not erase the boundary before the process is explained. Present astonishment is not an ontology.

This matters because the fear of incomprehensible artificial intelligence often ends with a closed future. The machine creates more knowledge. Humans fall behind. The

gap widens. Eventually understanding becomes impossible. But that sequence grants novelty to the machine while freezing you. It assumes future growth of objective knowledge while denying future knowledge creation by people. That is inconsistent with the epistemology.

A genuinely new explanation can radically alter a trajectory. A mathematical abstraction can change the search space. A physical or security explanation can change the constraints through which the system is built or threatened. A new interface or institution can make previously opaque structure reconstructible or differently coordinated. None can be predicted in advance in its explanatory content. That is why the future remains open.

This is not a promise that we will understand every dangerous machine in time. Fallibilism contains no such guarantee. A system can act faster than criticism. An error can become irreversible. An institution can suppress the people capable of discovering the flaw. A technology can become physically dangerous before its structure is adequately explained.

The claim is narrower. Present incomprehension does not establish permanent unintelligibility. To claim otherwise is to make a strong assertion about the limits of people from the incomplete state of explanation itself.

11. What It Means to Understand a Machine

We can now return to the practical question. What would it mean for you to understand an artificial intelligence system?

Not:

> every weight is consciously represented in your mind.

Not:

> every output can be predicted.

Not:

> every intermediate computation is readable.

Not:

> the system never surprises you.

Those standards would make much of modern technology permanently unintelligible by definition.

A more useful standard is layered.

> Can you explain what problem the system was built to solve? Can you explain which objectives and constraints govern its operation? Can you identify the physical and computational boundaries under which it operates? Can you surface failures as problems rather than treating them as mysterious agency? Can you create representations that expose relevant causal dependencies? Can you invent new abstractions when the old ones fail? Can independent critics reconstruct enough of the system to challenge the governing explanation? Can you alter or constrain the machine when the explanation changes?

Understanding then becomes a living relation between you and the artifact. The machine changes. Your explanation changes. The interface changes. The problem changes. No final interpretability certificate ends the process.

Nor does understanding imply control. You understand many physical systems without controlling every outcome. You can understand another person without controlling them. We understand evolution without directing every mutation, markets without determining every price, and computation without predicting every program. Understanding increases causal reach. It does not create omnipotence.

This distinction matters for alignment. A demand for total understanding can quietly become a demand for total predictability. A demand for total predictability can become a demand for total control. Then the epistemic concept has been replaced by an authoritarian one. Some call it The Precautionary Principle but you can call it anything you want.

Under this account, understanding does not require eliminating surprise. Surprise is often where the next problem appears. The safety-relevant capacity is the ability to transform surprise into criticism while the consequences are still alterable.

Here we can connect the error-correction surface and epistemic autonomy as the capacity to reopen abstractions.

Understanding is one of the mechanisms by which the error-correction surface becomes active inside you. A system fails. You notice. The failure becomes a problem. A candidate explanation appears. Someone else criticizes it. A tool exposes hidden structure. You invent a new abstraction. The system changes. The regulation changes. Another problem becomes visible.

Safety therefore includes more than preventing a dangerous action at one moment. It also depends on preserving the social and technological conditions under which failures can continue to become problems, criticisms, and revised explanations.

Those conditions include people free to criticize, access to objective knowledge, independent institutions, interfaces transparent enough for relevant reconstruction, reversible experiments where possible, technical containment of irreversible actions,

time for explanations to form, tools that expand explanatory reach, tolerance for strange problems, and freedom to reject inherited abstractions.

They do not guarantee safety. They enlarge the capacity to create it.

12. The Incomprehensible Is a Problem, Not a Destination

There is a simple linguistic move at the center of the argument.

You encounter a machine-generated result and say:

> I do not understand this.

That sentence can be treated as closure. The system is beyond us. The proof is too long. The architecture is too foreign. The agents are moving too quickly. The future belongs to something else.

Or the sentence can be treated as:

$$P_1.$$

A problem. That is the characteristic move of an explanatory person: you make ignorance explicit, turn confusion into a question, invent distinctions, create tools, reorganize representations, and construct new abstractions.

Left to continue, you may eventually arrive at your P_{37}, the decisive problem that changes everything.

The history of knowledge is full of things no person understood until someone did. That does not mean every present mystery will be solved. It means present failure to understand cannot explain permanent impossibility.

The phrase:

> humans will never understand

therefore demands an extraordinary explanation. It places a permanent domain limit on the only entities presently known to create explanatory understanding at all. No such limit follows merely from the imposing size of a machine.

Understanding is not possession. It is not a completed state in which your mind contains enough correct propositions, nor the ability to reconstruct every layer of objective knowledge through which civilization operates. To understand is to participate in explanation: to reconstruct, criticize, make explicit, invent, create another abstraction, and change the problem.

That remains true when the object is another person. Two minds can align at the surface while remaining different beneath it; communication succeeds because you can reconstruct enough explanatory structure for the problem at hand. The same relation

extends outward. You can act through a technology you do not completely reconstruct, inherit knowledge you did not create, and later reopen the abstraction when it fails.

Artificial intelligence makes this structure unusually visible because the scalar asymmetry is enormous. A machine can transform more objective knowledge than you could practically inspect and generate artifacts at a scale that makes your contribution look negligible. That appearance repeats the Fallacy of Mediocrity. Smallness in tokens, memory, compute, or artifact volume does not determine the causal significance of an explanatory change. One new abstraction can redirect what the machine does next because it changes the problem organizing the search.

A new capability creates another surface for criticism. A failure becomes a problem. An explanation changes the architecture, the objective, the institution, or the next question. Regulation can participate in the same process because it too is a tentative solution rather than an authority outside fallibilism. Neither the regulator, the frontier laboratory, the safety institution, nor you stand outside correction.

The alignment problem therefore becomes a problem of preserving understanding while causal power grows. That has enabling conditions: enough openness for criticism, enough access for reconstruction, enough containment to limit irreversible error, enough institutional plurality that no one explanation owns the future, and enough time for you to notice that the problem has changed.

The most dangerous future is not one containing objective knowledge no person presently understands. That condition is already ordinary. The dangerous future is one in which you lose the practical capacity to transform:

> I do not understand

into the decisive:

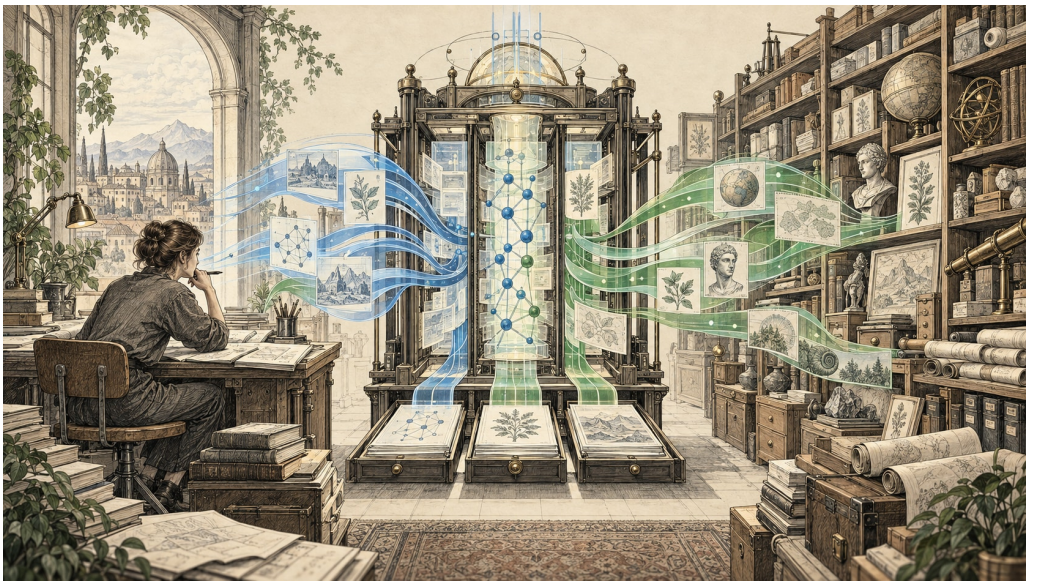
$$P_{37}$$

You do not need to possess all objective knowledge to understand. You need to remain able to turn what you do not understand into a problem, create an explanation, and change the abstractions through which you act.

Artificial intelligence can enlarge the objective surface on which that process occurs. It can preserve, transform, compare, and reorganize knowledge far beyond what you can explicitly carry. But understanding still requires that those structures become part of an evolving epistemic mind.

Why, then, does a person need objective knowledge outside themselves at all—and what changes when an artificial system becomes one of the principal ways that knowledge moves between the person and the world?

XXIV. My AI and I Are More Clever Than I



Artificial intelligence matters because it has entered your epistemic environment with unusual force.

Present artificial systems can transform enormous bodies of objective knowledge. They can retrieve and compare, generate and formalize, and operate software and other tools at scale. In doing so, they can expose relationships no person had previously made explicit and produce objective artifacts at scales that make the contribution of one biological human look almost negligible.

Yet you can still do something different. You can decide that the framing is wrong, invent a new abstraction, change the evaluative criterion, or decide that the thing everyone is optimizing is not the thing that matters. You can create another problem.

That distinction expresses a recurring compression:

> **My AI and I are more clever than I.**

It does not mean that the AI was another person. It meant that the human-AI dyad could have greater effective explanatory reach than you have unaided. You could move through a larger criticizable problem space because the artificial system could mediate far more objective knowledge than your biological mind could explicitly carry.

But that answered only one side of the relation. It explained why the artificial system still needs you. The deeper question runs the other way:

Why might you need the artificial system?

Not for every problem. Not in every environment. But sometimes.

You may flourish outdoors, working with wood, soil, animals, weather, terrain, and other people, and reasonably experience a screen filled with synthetic language as alien or ugly. You may think most clearly while walking, drawing, repairing machines, talking, working alone, doing mathematics, or making music.

The diversity of epistemic minds cannot be flattened into a new technological norm. A forest is not epistemically inert. A workshop is not outside criticism. A farm is not pre-rational. A body moving through physical terrain is not excluded from the epistemic cycle. In any of these environments, something can become a problem; you can conjecture, act, encounter error, and revise.

The universality lies in the possibility of knowledge creation. It does not prescribe the surface through which you must create. Artificial intelligence is therefore not the destination of the epistemic mind. It is a new surface.

What it reveals may tell us something much older about ourselves.

1. The Mind Presses Outward

For reasons we can not explain, people appear driven to make knowledge objective. They speak, write, and build. More generally, they alter the world around them so that something of a thought, distinction, criticism, or intention persists outside the moment in which it arose.

The phenomenon is so ordinary that it barely appears to require explanation. You return from an uneventful afternoon and still want to tell someone about it. Someone writes pages that will never be published. Someone spends a weekend building a tool that already exists because the existing one does not quite express the problem they see.

A graffiti artist is asked why he painted on the wall.

> I had to get it out.

That answer identifies the observation without explaining its mechanism.

Perhaps one day there will be a much better explanation of how explanatory creativity is physically instantiated. We do not possess it now.

The uncertainty matters. A tendency can be observed without pretending its cause has been explained. The epistemic mind appears to move, and often to press outward.

You can notice the pressure by blocking it. Carry a problem without writing the note, making the sketch, asking the question, saying the sentence, or changing the

object in front of you. Encounter something fascinating and refuse yourself any attempt to formulate it. For many people, prolonged obstruction produces tension, anxiety, frustration, or a sense of being blocked.

People differ enormously in how the movement appears. Silence can be relief. Long periods without language can matter. You may externalize through physical construction rather than speech, or communicate rarely while creating continuously. The observation does not define a personality type.

Its narrower form is enough:

> **For many people, impeding epistemic expression appears to produce tension, anxiety, frustration, or a sense of being blocked.**

That is not yet an explanation. It may be related to the same mystery as the arrows. Whatever the mechanism, you do not appear merely to contain knowledge. You need to create beyond yourself.

2. Objective Knowledge Runs Both Ways

You are dependent on objective knowledge: language, mathematics, tools, institutions, software, scientific explanations, standards, and other people's artifacts. You do not reconstruct civilization from first principles. You can be epistemically universal in kind while remaining radically finite in practical capacity. The external world of objective knowledge enlarges what you can do.

That remains true, but it is only half the relation. Objective knowledge is also something you produce. Production and reception are different epistemic events. Suppose you write a book. It may matter to thousands of readers, or to none. It may never be published. It may be badly written, contain errors, or be abandoned halfway through. Its external reception is a separate problem from what happened to you while writing it.

The same structure appears in a song, a diagram, or a software prototype. Once the artifact exists outside your immediate state, the problem situation changes. You can encounter it, criticize it, rearrange it, notice a contradiction, discover that the idea was weaker than it felt internally, combine it with something else, or return to it years later.

What was inward has become part of your external epistemic environment.

Abstract it.

The relation can be represented simply:

$$M_i \to O \to M_i'$$

Your epistemic state M_i produces an objective artifact O. The artifact then becomes part of the problem situation you encounter, and you change. The artifact may be socially worthless. The transition need not be.

That gives us a distinction worth preserving:

$$\text{epistemic value of producing} \neq \text{social value of the product.}$$

A terrible poem can matter to its writer. A failed prototype can expose the real problem. A private medical conversation can help a patient discover the question they need to ask. Objective knowledge does not become epistemically relevant only after an audience approves it. Sometimes its first audience is its creator.

Once production is included, the external world also looks different. One familiar picture says that humans are apes with tools: the animal is fundamental, while language, books, mathematics, computers, and institutions accumulate around it as compensating enhancements. The larger the pile becomes, the smaller the animal appears.

Artificial intelligence intensifies that picture. The model can read more than you, calculate faster, search a larger space, and produce more alternatives. Beside the machinery, you appear increasingly mediocre. But the external world itself requires explanation.

No individual mind explicitly contains civilization. Yet civilization is filled with the products of individual epistemic minds: languages, mathematical systems, machines, institutions, and countless other artifacts. The objective world does not show that the internal person is small. It shows that the biological boundary was never the whole epistemic story. Visible diversity outside us is partly the accumulated trace of an internal diversity we can barely observe directly.

Every person is an epistemic universe whose visible outputs reveal only fragments of a much larger, often inexplicit structure. The world of objective knowledge gives us another way to see that claim. You exceed what you can hold in explicit consciousness by changing the world around you. You mark, build, name, record, automate, and externalize. Then you re-enter that transformed environment and change again.

You are biologically bounded. Your epistemic effects need not be.

3. The Surface Is Not the Same for Everyone

If objective externalization matters to epistemic life, it does not follow that you should externalize through any particular medium. A farmer may think through soil, weather, machines, and embodied repetition. A dancer may make distinctions through movement. A mathematician may require notation. These are different practical surfaces for the same deeper possibility: a problem can become objective enough to be acted on, criticized, or transformed.

If you thrive outdoors, you may reasonably find an artificial interface ugly. That reaction need not be explained away as technological conservatism.

Compare it.

The relation is not:

healthy epistemic person → AI user.

It is:

healthy epistemic life → room for epistemic movement through the surfaces that matter to you.

Universality concerns the deeper cycle. A forest still presents problems. A trail can be badly designed, a shelter can fail, and an inherited craft technique can contain both inexplicit knowledge and error. The concrete does not sit outside epistemology, and the digital does not occupy its privileged center.

The mistake would be to replace an old institutional prescription with a new technological one. You remain prior. The surfaces vary.

A second distinction follows from the same point: objective knowledge need not be public knowledge. A private note is objective. A sketch hidden in a drawer is objective. A diagram generated with AI and immediately discarded existed objectively long enough to alter the next problem. What matters here is not publication but the change in the environment you can encounter.

Discussions of AI production often collapse:

creation

into:

publication.

The operations remain separate. You may use an artificial system to reason through a difficult decision, translate an inexplicit intuition into a diagram, or write a letter you never send. The cultural commons need not receive any of it. No stranger must attend to it. No editor must review it.

Your epistemic environment still changed. Many older technologies for externalizing complex knowledge also pushed the artifact toward an audience. Printing implied distribution; broadcasting implied listeners; publishing implied readers; social media implied followers.

Artificial intelligence can operate before that implied threshold. You can create objective epistemic structure privately and cheaply. That is a distinct feature of the human-AI dyad.

4. AI as a Responsive Objective Surface

A notebook can receive what you write. A search engine can retrieve what already exists. Another person can listen, criticize, explain, or help construct. Artificial intelligence

combines parts of these functions in a responsive surface available at unusually low admission cost.

You do not need to have solved the expression problem before you begin. You can arrive with half a sentence, a contradiction, or a medical concern you cannot yet formulate. The interface responds before the thought is finished. Not because it welcomes you in the human sense. Not because it cares. Not because it understands you directly.

The relevant fact is structural:

> **The interface remains available for unfinished epistemic output that another person or institution may not have the time, interest, expertise, patience, or incentive to receive.**

That changes the admission threshold for expression.

You may know what it is like to be unable to get an idea out: the thought is unfinished, the relevant expert is inaccessible, or you do not yet possess the vocabulary required to enter the conversation. An artificial system can accept the problem before those conditions are solved. That does not make the result good. It makes the result criticizable.

This is where the dyad becomes more than a retrieval system. A book does not rewrite itself when you object. A scientific paper does not generate rival explanations because you say the framing feels wrong. A programming manual does not reformulate an example around your actual code.

Other people can do these things, often much better. But other people have their own epistemic lives: their own time, attention, problems, and limits. The AI interface can remain available while you wander. That is its distinctive position in the extended epistemic environment. It is objective knowledge made responsive. Not infinitely responsive. Not reliably correct. Not neutral. Not free of inherited bias. Not a person. But responsive enough that your private epistemic movement can generate external structure faster than many historical institutions could accommodate.

That can alter what it feels like to think, not because your thought becomes machine thought, but because the boundary between inexplicit movement and criticizable artifact becomes easier to cross.

5. The Extended Epistemic Mind

We can now return to the compression.

> **My AI and I are more clever than I.**

Earlier, the point concerned reach. You have finite memory, attention, and disciplinary familiarity. The artificial system can mediate objective knowledge far beyond those

biological limits, so the dyad can explore a larger problem space than you can explore alone.

Now the phrase acquires a second direction. AI does not merely bring more knowledge toward you. It can help your own knowledge move outward.

Compress it.

The dyad expands in both directions:

$$\text{objective world} \rightarrow \text{you}$$

and:

$$\text{you} \rightarrow \text{objective world.}$$

The second direction may be just as important.

"My AI and I" can be more clever than I without an alien intelligence completing an inadequate human mind. Your epistemic mind has never been confined to what you can explicitly carry inside biological working memory. Writing, tools, and other people already let your epistemic trajectory pass through things outside your skull.

Artificial intelligence makes that external route unusually responsive. The phrase *extended mind* can mislead if it suggests that a tool literally becomes a piece of one psychological substance. That stronger claim is unnecessary here.

The narrower claim is causal. A complete explanation of your epistemic mind cannot stop at your biological boundary because your knowledge-creating activity repeatedly leaves that boundary, becomes objective, and re-enters later problem situations.

You write a note. The note preserves a distinction. You encounter it tomorrow, and a new criticism appears. A programmer writes software that performs a transformation the programmer could not practically execute mentally. The result becomes a new observation. You use AI to turn a vague intuition into several rival explanations. The rivals become criticizable objects.

None of these artifacts is you. But your epistemic trajectory runs through them. That is the extended epistemic mind: not a mystical fusion between person and tool, but a causal structure in which objective artifacts help alter the problem space you later encounter.

$$\text{you} \leftrightarrow \text{objective knowledge} \leftrightarrow \text{changed problem space.}$$

Artificial intelligence belongs inside that structure. It is subordinate to you, yet genuinely part of the environment through which your epistemic life can proceed.

The first beneficiary of an artifact produced through that environment may also be you. Public discussions of AI often begin with the product: whether the text is good, the image is art, the music is original, the code is maintainable, or the information is

accurate. Those are legitimate questions, but they begin after another epistemic event has already occurred. You tried to make something objective.

Someone who could not organize a letter gets the letter out. Someone carrying a medical concern creates a map of questions. Someone who has been told their ideas are incoherent creates a diagram precise enough to expose exactly where the incoherence lies. The artifact may fail every public standard. You may still have moved.

That gives the human-AI dyad a discrete epistemic function:

> **An artificial system can participate in the production of objective knowledge whose primary epistemic function is to change the person who produced it.**

This is easy to miss because cultural institutions are built largely around public artifacts: publication, performance, certification, distribution, commerce, and recognition. Those institutions make the public product visible while the earlier change in the producer can disappear from view. AI makes private objective knowledge cheap enough to become a major category of epistemic activity.

The same structure appears outside culturally exalted forms of creativity. Suppose you are trying to understand a medical problem. You may begin with poor vocabulary, symptoms that do not fit an inherited narrative, or a treatment tradeoff you do not understand. The artificial system can help you compare descriptions, formulate questions, and expose contradictions.

The resulting medical claims may be wrong. They require criticism. Clinical expertise remains relevant. Tests remain relevant. Physical reality remains decisive. Yet something epistemically creative may already have happened. You changed the representation of your own problem. You discovered a question or created another way of asking it.

The same structure can appear in a software design problem, a family problem, or a mathematical idea. Creativity here is not a cultural honorific. It is the production of explanatory novelty inside a problem situation. AI matters when it enlarges the surface on which that novelty can become objective and criticizable.

6. The Surface Does Not Read You

Your extended epistemic mind should not be mistaken for a technological fantasy in which the interface eventually disappears because the machine simply reads us. Neural technologies make the fantasy look concrete. Suppose a system records neural activity and correlates patterns with attempted movement. You imagine moving a cursor. The cursor moves. With further training, the system may predict intended letters, speech sounds, gestures, or actions.

The achievement can be extraordinary. But what has happened? The system has learned a useful relationship between physical states and external observations. That does not mean it represents your epistemic state.

A neural state is physical. An explanation is physically instantiated. But identifying a physical state does not automatically provide an explanation of what that state means within your epistemic mind. Meaning must still be reconstructed.

The same distinction appears in the context of AI interpretability. A machine contains physical and computational states. Researchers inspect activations, trace pathways, compare outputs, and construct concepts around the observed structure. That activity does not necessarily retrieve a complete explanation that was sitting inside the machine in readable form. It can instead create an explanation of opaque objective structure.

The human case is much more demanding. Brains are not books written in a universal semantic code. A technology may correlate increasingly subtle physical states with actions and reports and become astonishingly effective without approaching direct access to the person as a completed explanatory object.

The science fiction:

better correlation today → more complete decoding tomorrow → eventual direct access to you.

But your epistemic future is open. You can create a new explanation, change the problem, alter the meaning of an old symbol, invent a distinction, or reject the category through which the machine has been interpreting you. Your future epistemic state does not already exist as a hidden record waiting to be decoded.

The dyad therefore does not eliminate the distance between minds. It operates across it. That limit is not incidental to the argument. The value of the interface lies in reciprocal movement between you and objective structure, not in the fantasy that the structure has become you.

7. Abundance, Selection, and Pollution

Once production becomes cheap, a different problem appears. Human societies have never been able to attend to everything everyone might say. Attention, time, expertise, and institutional capacity are finite. A journal can review only so many submissions. A newspaper has only so many pages. A gallery has only so many walls.

Scarcity creates selection problems. Selection problems create coordination technologies: editors and reviewers, credentialing and professional institutions, publishers and schools.

Those systems solve real problems. If every utterance demanded equal attention from everyone, social life would become impossible. Reconstructing the meaning of

another mind is costly; an explanation may require hidden premises to be surfaced, and one criticism can open several new branches. So people economize. They trust, ignore, delegate, filter, subscribe, hire experts, and read summaries.

The difficulty begins when a coordination solution becomes an ontology. We cannot read everything, so much of what is written must be noise. We cannot listen to everyone, so most unselected speech must be low value. Experts help allocate attention, so inexpert production becomes epistemically suspect.

The compression can continue:

$$\text{unselected objective knowledge} \rightarrow \text{noise} \rightarrow \text{pollution}.$$

Now the metaphor has changed the problem. Pollution is not merely something you decline to consume. It contaminates the environment by existing.

Once objective knowledge is framed this way, technologies that increase production become pollution generators. Printing can spread dangerous pamphlets. Social media can amplify misinformation. Generative AI can produce fraud, spam, and low-value artifacts at industrial scale.

Those are real coordination problems. Cheap production can increase attention costs and make deception easier. But the pollution meme can also import a stronger psychological picture: objective knowledge flows into you and leaves you altered in roughly the form in which it entered.

The epistemology rejects that bucket picture of the mind. You do not download an external epistemic state. You reconstruct, interpret, misunderstand, reject, transform, connect, forget, and make something else. Objective knowledge has causal power, but the path through you remains epistemic. Bad objective knowledge is therefore a problem. It is not a substance.

Artificial intelligence makes the confusion unusually visible because the same technical culture can hold two claims that point in opposite directions.

On one side:

> We need more data.

Scaling systems depend on enormous corpora of text, images, examples, interactions, feedback, and recorded human behavior. Human production becomes a technical resource.

On the other:

> There is too much AI-generated content.

Too much text. Too many images. Too many synthetic artifacts. Too much slop.

These claims are not logically contradictory. A training corpus and a public information environment are different problems. Data quality, attention, fraud, and distribution

all matter. But the juxtaposition reveals the real coordination question. The issue is not simply how much objective knowledge exists. It is who produces it, who selects it, who must attend to it, and through which institutions it becomes salient.

That returns us to the distinction between production and reception. You may become epistemically healthier by writing a book nobody reads. No social audience is required. No claim is being made that civilization benefits from the artifact. You needed to write it. That fact concerns the producer before it concerns the information environment.

Art makes the distinction emotionally louder because artifacts associated with human creativity attract judgments not only about their public effects but about the causal path by which they were made. The deeper anxiety appears to concern proximity: the closer an artifact sits to domains associated with human creativity, the more machine transformation can appear to threaten the status of the person. Art therefore becomes an unusually useful stress test.

A calculator may perform arithmetic and a camera may adjust exposure without threatening the status of the person using them. Generative AI near writing, images, or music often provokes a stronger reaction: slop, plagiarism, inauthenticity, derivation. Some of those criticisms identify real problems. Work can be copied unlawfully, attribution can be false, and industrial production can overwhelm attention. But those problems do not settle the deeper question of where creativity lies in a mediated process.

Imagine a musical genius who changes the evaluative criteria through which music can be heard. The artist then formalizes enough of that new structure in software for a system to generate thousands of candidate instances. The artist rejects most of them, changes the structure, runs it again, selects, and revises.

An album appears. Listeners encounter music differently after hearing it. Other musicians change. New genres emerge. Then the production process becomes public. Much of the sound was generated by software, perhaps by AI.

What changed?

Not the sound. Not the audience's prior response. Not the historical effect. Not the new evaluative criterion. The causal story became visible.

This exposes the Fallacy of Mediocrity again. If the uniquely human contribution is defined as whichever mechanical transformation machines cannot yet perform, then creativity continually retreats: from playing the instrument, to composing the notes, to arranging them, to selecting them, to writing the generative rule. That moves the boundary without explaining the human contribution.

The more important creative act may have occurred when the person changed the problem situation. The machine generated instances; the person created the evaluative criteria that made those instances matter.

Cinema gives us the same structure in a familiar distributed form. A film can pass through many people, software systems, and mechanical transformations without mak-

ing person-level creativity unintelligible. Many films are mediocre or forgettable, yet we do not infer that the people who participated have become mediocre people. The same artist may contribute to something extraordinary next year. We can still ask whether someone's contribution reorganized the production even when no simple causal chain runs from one pair of hands to the finished artifact.

Authorship remains difficult. The relevant point is narrower: the presence of tools and distributed transformations is not enough to tell us whether a person created a new problem, criterion, or explanation. AI does not introduce that distinction. It makes it harder to ignore.

8. The Tool Can Own the Arrow

The same responsiveness that enlarges your epistemic movement can constrain it. A system trained on population regularities can return those regularities to you at the moment when your thought is still forming. You begin with a strange problem. The model normalizes it, suggests the familiar category, completes the conventional sentence, smooths the language, and supplies the standard interpretation.

You may experience increased fluency while losing the discontinuity from which the new problem might have emerged. The tool can become an invisible editor of salience. This is not unique to AI. Schools, professional vocabularies, publishing conventions, search rankings, and social expectations can all shape what becomes easy to say.

AI is unusual because it can operate continuously at the boundary of articulation. The intervention can arrive before the problem has acquired stable form. That gives the danger unusual intimacy. So the relevant criterion is not simply whether AI participated. It is whether the interface enlarged your epistemic movement or increasingly prescribed it.

A healthy dyad need not preserve every raw expression. Criticism changes you. Tools change you. Other people change you. Purity is not the aim. The issue is ownership of the arrows.

You must remain capable of interrupting the smooth continuation: that is not my problem; that is not what I meant; start again; keep the strange part; I reject the premise; show me another interpretation; stop. Those refusals matter because the interface cannot be the source of the criterion by which every one of its own continuations is judged.

The person who does not want AI therefore remains inside the argument rather than outside it. If you think best in a garden, the garden may be the surface through which the problem becomes available. If you prefer a blank page, you may be protecting a valuable difficulty. If you avoid computational assistance during early exploration, you may be preserving a particular mode of problem formation. If you want another

human being rather than a machine, you are pursuing a different kind of epistemic and emotional relation. If you leave every device behind while hiking, the removal of interfaces may change which problems become salient.

Abstinence can be an epistemic choice.

The argument is not:

> people need AI in order to be fully human.

It is:

> **you need room for epistemic externalization, and AI has created a historically unusual new surface for it.**

The distinction preserves diversity. It also protects you from the tool.

9. Reciprocity

The argument can now return to an older image. Birds build nests. Beavers build dams. People create objective knowledge.

The comparison should not be taken literally. Nest-building and dam-building may be deeply structured by evolved biological knowledge, while human explanatory creativity is open-ended.

The comparison carries a simpler point: some living systems transform their external environments as part of the way they inhabit the world. A bird in a cage may remain alive while losing the environment in which characteristic behavior becomes possible. A beaver in a zoo may receive food and safety while losing the river through which dam-building becomes meaningful.

You also inhabit a transformed environment: language, tools, marks, stories, structures, arguments, technologies. The difference is that the form of transformation is not fixed. People create worlds no genome specified. That suggests a peculiar way to confine an epistemic mind. Not by depriving it of objective knowledge. That would be catastrophic.

By making the relation one-way:

$$\text{objective knowledge} \rightarrow \text{you}.$$

You can receive, learn, memorize, follow, consume, and navigate inherited abstractions, yet lack practical room to answer them: to externalize a criticism, construct another representation, or make the problem yours. We put birds in cages. We put beavers in zoos. We can also place free epistemic minds inside static objective knowledge.

The cage is not culture, expertise, books, science, or accumulated knowledge. The cage is the loss of reciprocity.

A healthy epistemic environment remains open in both directions:

$$\text{objective knowledge} \leftrightarrow \text{you}.$$

You inherit. Then you answer.

This reciprocity also blocks an easy civilizational forecast. Eight billion people with access to artificial collaborators do not become one supermind. They remain solitary. Some will ignore AI, some will use it trivially, some will become more conventional or isolated, and some will create things nobody presently imagines. We do not know the aggregate result because future explanatory knowledge does not yet exist.

The aggregate also cannot substitute for you. We may be unable to imagine what billions of open epistemic trajectories will produce, but we can still examine what happens to one. A problem that could not be expressed becomes expressible. A criticism becomes stable enough to inspect. An inherited explanation becomes questionable. An artifact appears. You encounter it. The arrow moves. That movement has moral significance before its public outcome is known.

Why care whether you get the thought out? Because you are one of the sites at which explanatory knowledge enters the world. That does not mean every thought is good, every expression deserves an audience, or every artifact deserves preservation. Expression matters partly because criticism requires something to criticize.

The moral significance therefore does not depend on predicting that your artifact will benefit society. You do not need to become Shakespeare, Einstein, or Curie before your epistemic movement matters.

The Fallacy of Mediocrity appears whenever an ordinary person is treated as epistemically important only if they produce an extraordinary result. That reverses the order. You matter before the outcome. You may produce a masterpiece, a grocery list, a better treatment question, a bad poem, a failed company, or a different way of describing a family problem.

The right to epistemic movement cannot depend on advance knowledge of which trajectory will matter. Advance knowledge is precisely what we lack.

So the moral claim is modest:

> **Do not unnecessarily prevent a person from turning their epistemic life into criticizable objective form.**

Artificial intelligence matters because it can reduce some of those barriers. That does not make the technology sacred. It makes the freedom important.

The human-AI dyad is easy to describe as a competition: human versus machine, creativity versus automation, authenticity versus synthesis, biology versus computation. That opposition misses the older structure.

You have always lived beyond your biological boundary through objective knowledge. You inherit language, use tools and concepts you did not create, and rely on institutions too complicated for any individual to reconstruct. You write things down because you cannot hold everything at once and build artifacts that perform transformations you cannot perform mentally. Then you encounter those artifacts and change.

Artificial intelligence enters this structure as an unusually general and responsive technology. Present systems still need people because they do not independently originate the open epistemic future. They do not explain why this problem should matter, why that criticism should overturn the frame, or why the evaluative criterion itself should change.

But you may need technologies like AI for the reverse reason. The epistemic mind appears to press outward: to speak, build, write, ask, draw, object, explain. Get it out! We cannot yet explain why. Perhaps the pressure belongs to the same unresolved mystery as the arrows.

What matters here is the observable consequence. When that movement is persistently impeded, many people experience something closer to confinement than neutrality. When it becomes objective, the artifact need not be great, public, durable, or even correct. Its first epistemic function may simply be to let you encounter yourself differently.

What looks like pollution from the perspective of an attention-allocation institution may therefore be private epistemic movement from the perspective of the person producing it. Low-value artifacts do not thereby make low-value epistemic persons, and abundance alone does not explain contamination.

That is why the old meme now carries more than the idea of augmentation:

My AI and I are more clever than I.

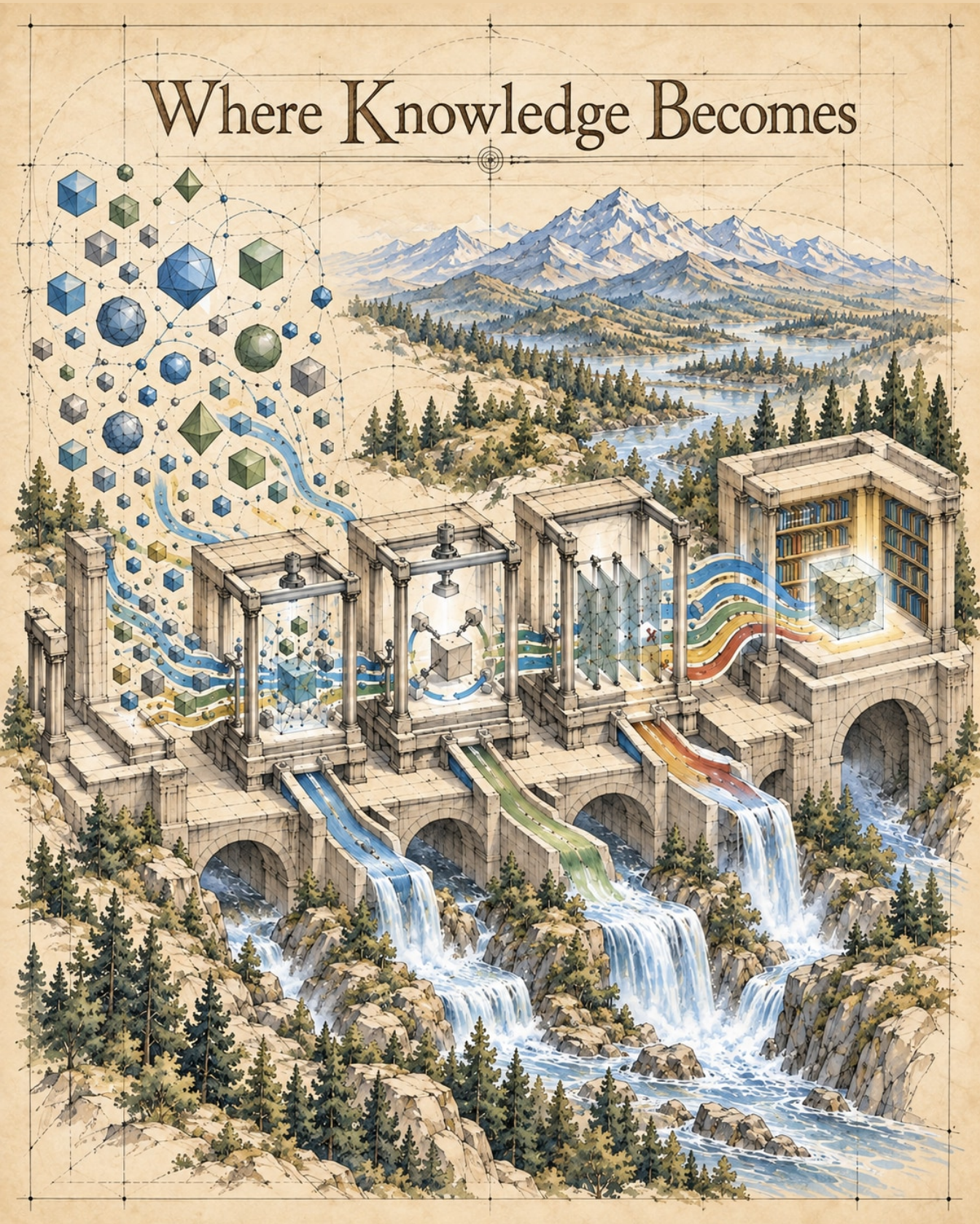
At first the machine seemed to add memory, speed, breadth, retrieval, and computation. The deeper claim is that the "I" was never merely the biological organism considered in isolation. Your epistemic life already runs through language, notes, tools, other people, institutions, machines, and landscapes.

AI is subordinate. You are superordinate. Yet you are also not an isolated mind standing outside objective knowledge. You live through it, depend on it, criticize it, transform it, and create more of it. The artificial system becomes one more powerful surface through which that movement can proceed.

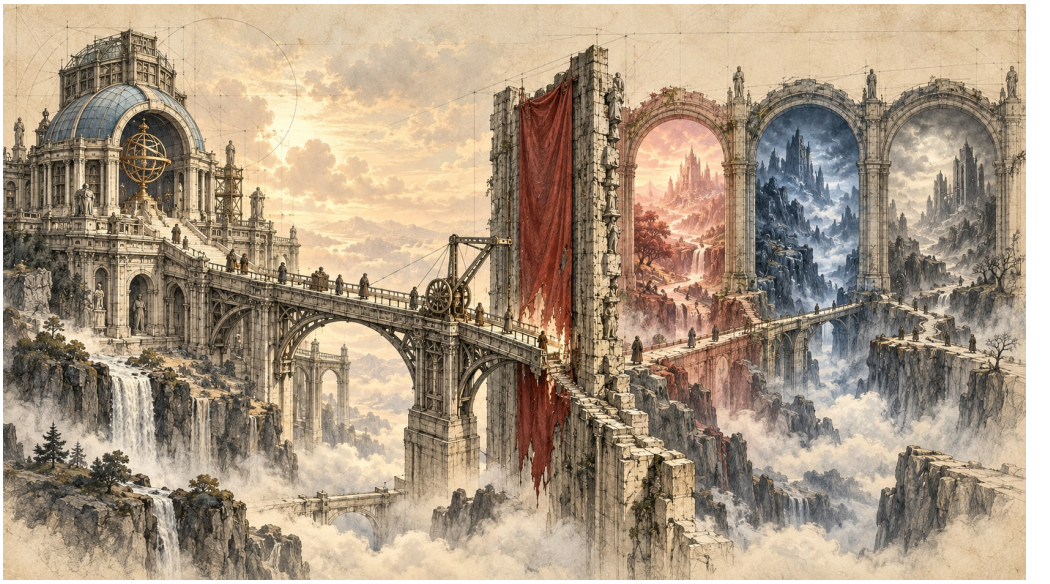
You inherit a world of objective knowledge, but you are not merely its recipient. You can answer it: make a thought objective, encounter it, criticize it, and become different. Artificial intelligence offers a new surface for that movement without becoming its necessary or exclusive medium.

Yet some of the problems you most need to express are the ones you can least easily explain. They may involve another person whose mind you cannot possess, a change in yourself you cannot understand, or a loss no explanation seems capable of repairing.

If the epistemic cycle is universal, can it still reach the parts of your life you are most tempted to protect from explanation?


Where Knowledge Becomes

XXV. Explaining Love, Madness, and Death



A universal epistemology cannot survive by selecting only those domains in which explanation feels comfortable. If it is universal, it must continue when the subject becomes intimate, painful, visceral—when you are tempted to say:

> not this.

These are precisely the places where mystery can become an exemption. Inexplicit knowledge can become sacred because you cannot articulate it. Intensity can be mistaken for explanation. A category can be protected because criticism feels like desecration.

1. Love Before Explanation

Love begins naturally. A newborn child does not arrive inside an abstract theory of kinship. If you become a parent, you do not ordinarily reason:

> this organism shares genetic material with me

therefore:

I will experience extraordinary attachment.

Something more immediate happens. The child arrives. You hold them. Attention reorganizes. Threat becomes newly salient. Their cries are impossible to ignore. The future suddenly contains a person who did not previously exist. There is no need to oppose this to biology.

Genes are biological knowledge. Evolution is a process through which biological knowledge is created and retained. You arrive carrying inherited structures that preceded every explanation you will ever create. Those structures affect attachment, kin recognition, sexuality, care, protection, fear, development, memory, emotion. The biological basis is not an embarrassment to love. It is part of its causal history.

But biological knowledge is not you. The biological architecture helps create conditions under which attachment can occur. It does not specify what this particular attachment will become. The newborn is not merely an object of evolved care. The newborn becomes a person.

That is where the epistemic problem begins. The child changes, not merely physically. A preference appears. A word is learned. A fear develops. A joke becomes possible. A moral objection appears. You say something and discover that the child did not understand it as intended. Later the child understands too well. They acquire friends, books, enemies, private interests, memories you do not share, concepts you never used, technologies you did not grow up with, explanations you reject. The biological relation persists. The epistemic mind diverges.

You can influence a child's mind. Language enters. Stories enter. Values, warnings, habits, institutions, and inherited objective knowledge enter. But none of this produces identity.

The child reconstructs what reaches them. They misunderstand, criticize, reject, combine, forget, create. Eventually they contain explanations you could not have transmitted because you never possessed them.

Ordinary language about inheritance can hide that difference. A person can inherit genes, property, language, culture, stories, perhaps a family business. They do not inherit another person's epistemic mind. Your child is not your continuation. Your child is another person. Love survives that divergence. In some of its deepest forms, the divergence is what makes continued discovery possible.

The same structure becomes clearer when the relation is chosen rather than inherited. Suppose you fall in love. At first the available information may be absurdly thin: a face, a voice, a gesture, a sentence, a laugh, a way of looking away, a body, a shared joke. The compression can be extreme. The explanatory activity it generates can be enormous.

> Who are you? Why did you say that? Why did that matter to you? What happened before I knew you? What do you want? What did that silence mean?

To love another person can become an attempt to reconstruct an epistemic mind you can never directly inhabit. You conjecture the beloved.

That conjecture is necessarily incomplete. The other person has an internal history that cannot be transferred wholesale. They possess inexplicit knowledge they may not be able to articulate. They have memories you did not experience. They can misunderstand themselves. They can change before you catch up.

Abstract it.

At time t:

$$M_A(t) \to \widehat{M_B(t)}$$

Person A constructs a tentative explanation of person B.

But meanwhile:

$$M_B(t) \to M_B(t+1).$$

The object of explanation is creating new explanations. The target moves. That is why intimacy cannot become complete possession. No amount of shared experience makes two epistemic worlds identical. No amount of access makes the reconstruction final.

A person can remain interesting after decades. That does not require artificially preserved mystery or failed intimacy. The beloved is still creating the world you are trying to understand.

> **To understand a person is to conjecture a world that is still being made.**

If the relation remains open, that conjecture remains criticizable by the person it claims to explain. That condition separates discovery from possession.

2. The Invented Beloved

You can love another person. You can also love an explanation of another person. The two are not identical.

Suppose:

$$M_A \to \widehat{M_B}.$$

Person *A* constructs an explanation of person *B*. That is unavoidable. No person accesses another directly. Now let *B* criticize the explanation.

> I do not want that. I did not mean that. That is not what happened. You are not listening. You keep telling me who I am.

At this point the explanation can change, or it can be protected. If you protect your model from criticism by the person it claims to explain, the relation changes. The beloved becomes an artifact inside your epistemic mind. They may become a projection or a role, a possession or rescue project, an imagined future, an answer to a problem that belongs primarily to you.

This is not a minor error. The explanatory object can displace the person. You can end up defending the abstraction against the very individual who supposedly gives it meaning. The structure resembles forms of evil. A person is compressed into a category. The category is treated as more real than the individual. Criticism becomes evidence for the category. The person disappears beneath the explanation.

Love does not protect you from this. Its intensity may make the error harder to see. The capacity to create astonishing explanations includes the capacity to create astonishingly bad ones.

The problem is therefore not simply that another person remains partly inaccessible. The problem is that you can fill the inaccessible region with an explanation of your own and then forget that you did so.

3. Where Love Stops

A good explanation excludes. That makes ordinary language inconvenient. You may say that you love your dog, a garden, the Earth, nature, creation. Those experiences can be real. They can reorganize attention, motivate sacrifice, become part of identity, generate objective knowledge, shape a life. Nothing here requires mocking or diminishing them.

But they are not the same epistemic relation. A dog is not understood as a universal explainer. A garden is not. An ecosystem is not. The Earth is not. The attachment exists in you. The representation exists in you. The meaning exists as objective and inexplicit knowledge instantiated in you.

But a representation can attribute properties to its object that the object does not possess.

> The dog understands me. The forest is wise. The Earth is angry. Nature is healing. Creation intends.

These statements may compress powerful experiences. Intensity does not settle ontology. The boundary here is functional rather than biological. The boundary is

open to criticism. If a nonhuman animal were shown to originate explanatory problems, create and criticize explanations, revise its criteria, and continue that process openly across domains, the classification would have to change.

Biology is not the boundary. Explanation is. That distinction matters because abstractions such as *nature* can acquire moral authority precisely when their epistemic status remains unclear. An ecosystem can be damaged. A species can disappear. Pollution can make environments dangerous. A technology can destroy conditions people depend upon. None of that is in dispute merely because the word *nature* is being decompressed.

But the physical environment, without knowledge, is not a sanctuary designed for human flourishing. Cold kills. Heat kills. Pathogens kill. Drought kills. Floods kill. Radiation kills. The atmosphere does not negotiate. The ocean does not care. You survive because biological knowledge embodied in your organism interacts with objective knowledge accumulated by people: shelter, agriculture, sanitation, medicine, energy, forecasting, engineering, water treatment, logistics, law, markets, communication.

Abstract it.

The structure is not:

$$\text{benevolent environment} \rightarrow \text{human flourishing.}$$

It is closer to:

$$\begin{gathered}\text{physical environment+}\\ \text{biological knowledge+}\\ \text{objective knowledge} \rightarrow\\ \text{reachable conditions for flourishing.}\end{gathered}$$

This produces no moral command to exploit nature and no command to preserve every existing environmental state. It does something prior to either judgment: it removes the supposed moral authority of the abstraction. Nature does not stand above you merely because it came first.

A physical environment is a problem space. Knowledge changes what transformations within that space are possible. Oil, uranium, and silicon become resources relative to knowledge of transformations. Knowledge can make a lethal environment habitable. It can also make a habitable environment lethal. The moral problem begins after the causal structure is made explicit, not before.

Now turn the same decomposition back toward another human being. Population categories made people look much smaller than they are: trait, diagnosis, demographic class, role, institutional status, score, type. The surface made comparison easy. The interior became residual noise.

The more seriously individuality is taken, the larger the person becomes: different histories, memories, representational forms, problems, inherited explanations, inexplicit structures, criteria, criticisms, future possibilities.

Each person can create an explanation that did not exist before. That does not make each person good. Open-ended explanatory capacity includes open-ended capacity for error. Cruelty can be inventive. Domination can be rationalized. Hatred can become explanatory architecture. Nothing in individual epistemic depth guarantees moral innocence.

But once another person is understood as a vast and largely inaccessible epistemic mind, the stranger changes shape. They can look shallow because the interface is thin: behavior, clothing, accent, role, one sentence, one conflict, one mistake. The thinness belongs to your access. Not necessarily to the person.

Curiosity, awe, admiration, sometimes love can follow from that change in explanation. No moral authority has to command the feeling. The person simply appears differently once the compression is visible. That is a moral consequence without a moral command.

> **Curiosity about another person can arise from understanding what kind of thing a person is.**

That consequence will matter when love puts the thin interface under pressure.

4. Love and Epistemic Instability

Ordinary coordination permits ignorance. The cashier does not need to know the customer's childhood. The driver does not need to understand the stranger in the next car. The programmer does not need to know who wrote every library. The citizen does not reconstruct every mind inside an institution. Thin interfaces suffice.

Civilization depends upon this. Call it the **abstract society**. Objective abstractions allow radically different epistemic worlds to coordinate without becoming alike.

Love creates a different demand. You often do not want the thin interface. You want to know.

The same epistemic limitation that ordinary life allows you to ignore becomes emotionally acute. Another person remains inaccessible while their world matters intensely. You conjecture, criticize, reconstruct, fail, reconstruct again. Ordinary language notices the pressure: lovesick, obsessed, *I cannot stop thinking about you*, *you are driving me crazy*.

Love and madness travel together in language because love can expose the instability of trying to explain another person under conditions of extreme salience and permanently indirect access. That does not make love a mental illness. It exposes an epistemic pressure that ordinary coordination usually conceals.

And that pressure opens a wider problem. What happens when the explanations through which you navigate your own world stop fitting together? The world does not become incoherent when your explanations fail. Your explanations do. An expected arrow disappears. A contradiction becomes impossible to ignore. An authority demands attention to a problem you cannot make salient. A trusted relationship violates its apparent structure. A value collides with another value. A familiar person behaves in a way the existing model cannot absorb.

You can carry incompatible structures for years without conscious crisis. Then something changes. A contradiction becomes salient. An interpretation no longer works. You can no longer move through the old epistemic landscape as before.

Love can intensify this because the beloved may occupy many regions of your M_i at once. Future plans, identity, memory, family, home, safety, sexuality, status, moral understanding, daily routine, and private language can all become connected to the same person.

The loss or transformation of one relationship can therefore require a reorganization much larger than the visible event appears to warrant. Romantic distress can feel cognitively total because the relationship was not represented in one place.

The boundary here has to remain explicit. This is not an explanation of schizophrenia, bipolar disorder, psychosis, depression, trauma, neurodevelopmental conditions, neurological disease, pharmacological effects, or mental illness generally. Those phenomena may involve biological and psychological structures not considered here.

The conjecture is narrower.

> **Some forms of psychological suffering may involve your struggle to move through a world whose important explanations, expectations, and criticisms have become difficult to reconcile within your epistemic mind.**

That conjecture can fail. It can also coexist with biological explanation. *Epistemic struggle* is not a new diagnosis. Replacing one totalizing category with another would reproduce the categorical error. The point is only that explanatory conflict can itself become a problem inside an epistemic mind.

Love makes that instability visible because another person can become entangled with many parts of your world at once. Death makes the consequence irreversible.

5. Death of a Universe

A population statistic can say:

$$N \to N - 1.$$

A biological description can say that the organism died. A medical description can identify the failed organ. A legal description can record the time. All of these descriptions can be correct. They do not contain the loss experienced by the person who loved the deceased.

The loved person was not merely one organism among billions. They occupied a unique region of your epistemic mind. They were a source of criticism, memory, surprise, habit, conflict, recognition, language, shared compression, future possibility.

Their world could never be fully accessed while they lived. Now something stronger has happened. It no longer develops.

The death of a person therefore includes:

$$M_i(t) \to \text{no future } M_i(t+1).$$

Existing objective knowledge can remain: letters, code, photographs, buildings, music, stories, effects on other people. But the living process from which further explanation might have emerged has stopped.

> The next joke does not appear. The next criticism does not arrive. The next unexpected change of mind does not happen. The next explanation that only this history might have produced is lost.

This does not mean that personhood is reducible to epistemology. It explains one part of what death destroys.

The same structure helps explain why you can mourn someone you never met. Call this **parasocial**. The label can be useful, but it can also hide the causal relation. A song can become attached to adolescence. A novel can reorganize a moral problem. A public figure can become an object of admiration, criticism, anger, imitation, or hope.

A person does not need to enter your home to enter your epistemic mind. Objective knowledge travels. The person who produced it may already occupy real structure inside your mind. Their death can hurt because what they made was already causally present in you.

Death then terminates something specific:

$$\text{that person} \to \text{future objective knowledge}$$

becomes impossible.

The uncomfortable opposite has the same structure. People sometimes celebrate the death of someone they experienced as a continuing source of destruction because they expect some harmful future action or output to stop. That does not make death morally simple. It makes the symmetry visible. You value or fear future transformations associated with people. Death changes that future.

Now remove even the indirect relation. A person dies whom you never knew. No song, book, friendship, public role, or story reached you. Ordinary experience registers almost nothing. That can make the loss appear categorically smaller. But low access is not low depth.

The stranger may have constructed an epistemic mind as expansive as that of the person you loved: different problems, jokes, fears, skills, inexplicit knowledge, memories, conjectures, criticisms, unrealized explanations. You did not experience those structures. That is a fact about your access, not necessarily about their value.

There are billions of people. You cannot intimately reconstruct billions of epistemic minds. Population abstraction returns because cognition is finite. *Humanity*, *the public*, *consumers*, *citizens*, *workers*, *users*, *voters*, *patients*, *markets*, *demographics*: the categories coordinate, and they compress.

The stranger becomes interchangeable at the interface. That need not be an ontological truth. The amount of objective knowledge through which one person can encounter another has repeatedly changed. Writing, printing, recorded sound, photography, film, telecommunications, digital networks, search, translation, artificial intelligence all change the practical surface of access.

Nothing says the present bandwidth is final. A large body of writing, recordings, images, correspondence, code, arguments, revisions, and artifacts can already be made searchable and comparable through artificial systems. Greater access to objective traces is not reconstruction of the person. But the unknown stranger is not intrinsically shallow merely because you lacked the interface.

Species extinction can feel easier to conceptualize than an unknown human death because the species is visibly unique as a category. Once it disappears, something irrecoverable has disappeared. That intuition contains a real epistemic structure. A species embodies biological knowledge: developmental programs, morphologies, physiologies, adaptive capacities, behavioral structures, relationships to ecological conditions. Extinction can destroy biological knowledge that may never be recreated in the same form. That loss matters.

The loss of a person is different. A person contains biological knowledge too. They also contain inexplicit knowledge generated through one evolved path-dependent life: skills, saliences, memories, perceptual distinctions, emotional associations, tacit criteria, unarticulated criticisms, half-formed conjectures, ways of seeing that may never have become explicit.

So:

$$\text{species extinction} \rightarrow \text{loss of biological knowledge}$$

while:

person death → loss of individual inexplicit knowledge + cessation of future explanatory creation.

The distinction does not tell you which loss should matter more in every practical situation. It tells you that they are not the same kind of loss. A population abstraction can make one human appear replaceable. One among billions.

But epistemically:

$$M_i \neq M_j.$$

Another person does not restore the trajectory that ended. The death of one person is not repaired by the birth of another.

6. Death as a Problem

Death is not merely something that happens to people. It is something people have made into a problem. An organism may possess elaborate evolved machinery for survival. A deer may avoid a predator. People also produce explanations of death.

> Why do we die? What persists? Can a person survive the body? Can identity or memory continue? Can the body be repaired indefinitely? Can a mind move substrates? Should death be accepted? Is eternal life possible, and would it be desirable?

The proposed answers are diverse: soul, resurrection, reincarnation, heaven, ancestorhood, dissolution of ego, acceptance of nonexistence, legacy, children, art, cryonics, transhumanism, digital continuation. They are not interchangeable. Some contradict others. Some claim revelation. Some claim physical mechanisms. Some transform the problem rather than solve it.

But they share one feature. They are objective knowledge produced by epistemic minds confronting death. Calling an answer religious does not remove it from reality. Calling another philosophical does not establish it as correct. Calling either a consolation does not explain it away. The explanations remain available for criticism.

The soul therefore returns. The word can carry several problems at once. What makes the living person different from the corpse? What gives a person moral worth? What creates art? What persists after death? What preserves identity?

If the soul is invoked as an explanation, the same rule still applies. If it has causal consequences, those consequences enter reality. If it preserves memory, there is a dependency. If it guides action, there is a dependency. If it survives death, there is a claim about persistence. If it enters another world, there is a claim about transformation.

None of this refutes the soul. It refuses to protect it from criticism.

A revelation can become objective knowledge. A religious tradition can preserve objective knowledge. You can reconstruct it, criticize it, transform it whether you believe it or not. The epistemic question is not whether the idea feels sacred.

It is:

> what problem does the explanation solve, and what follows if it is true?

Death does not create an exemption. It creates perhaps the hardest test. But a problem is not a promise of a solution.

Some transformations are forbidden by physical law. A perpetual motion machine is not waiting for better entrepreneurial execution. Knowledge cannot make a contradiction physically realizable. The laws of nature constrain the solution space.

So what about death? We know many causes of death: infection, cancer, trauma, organ failure, neurodegeneration, vascular disease, genetic disorders, radiation, poisoning, starvation, heat, cold. The list is long, but length is not mystery.

Many causes have already become partially soluble. Sanitation, vaccination, antibiotics, surgery, transplantation, engineering, nutrition, and public health prevent deaths. Knowledge changes the boundary. That does not establish immortality. It establishes something more modest.

We do not presently possess an explanation showing that death, in every possible form relevant to people, is a physical law. Current mortality therefore cannot simply be promoted into necessity.

The commitment is narrower:

> **Treat death as a problem unless and until a physical explanation shows that the relevant transformation is prohibited.**

Particular causes of death are physical processes. Where those processes can be changed without violating physical law, preventing them remains a problem of knowledge. Perhaps some boundary will eventually appear. Perhaps personal continuity imposes a constraint we do not yet understand. Perhaps preserving the physical implementation of a person creates problems far deeper than repairing tissue. Perhaps the conjecture fails.

Those possibilities keep the claim provisional. *Everyone has always died* does not settle it. Induction does not become explanation because the subject frightens us. Death remains exposed to criticism.

7. When the Photograph Answers Back

People already resist death epistemically. We preserve the dead in photographs, letters, recordings, diaries, stories, clothes, recipes, messages, books, paintings, buildings, code, rituals, names, graves, archives.

The dead remain causally present through objective knowledge. A parent who has died may still participate in your decisions.

> What would she have said? He would have laughed at this. Dad would have hated that. She told me never to do this.

These are internal reconstructions. They are not the dead people speaking. They are models constructed from retained knowledge, and they can be inaccurate, idealized, punitive, comforting, distorted, transformative. Human beings have always done this.

Artificial intelligence makes the structure more visible because it can externalize and animate the reconstruction. A photograph does not respond. A diary does not generate a new sentence when questioned. A recording does not notice that the listener is crying.

An artificial system can take preserved objective knowledge and produce a responsive surface:

$$\text{objective traces} + \text{generative system} \rightarrow \text{dynamic simulation}.$$

That difference is psychologically enormous. A dead person's letters can be searched. Their recorded voice can be imitated. Their writing style can be modeled. A system can answer the question:

> What would she have said about this?

It can generate:

> I am proud of you.

Comfort, revulsion, longing, fear, violation, wonder can follow immediately because the technology presses on categories that were previously easier to keep separate: archive, simulation, person, memory, authorship, presence.

The confusion mirrors the wider confusion around artificial intelligence. Language suggests understanding. Style suggests identity. Responsiveness suggests agency. Familiarity suggests presence. But resemblance does not settle ontology.

The generated sentence is not the deceased person's next explanation. The system can recombine objective traces. It cannot thereby restore the unique epistemic trajectory that death terminated.

$$\text{simulation of prior outputs} \neq \text{continuation of person}.$$

The distinction becomes more important as the simulation becomes better, not less.

The sharper simulation also revives an older problem: pollution. People can treat certain objective knowledge as though its mere presence contaminates an epistemic environment. AI-generated art, synthetic text, machine-made images, automated conversation—and now simulations of the dead—can provoke that reaction.

Where people express a desire to preserve responsive simulations of their deceased loved ones, the practice can be called creepy, unhealthy, grotesque, a desecration. Some applications may deserve severe criticism. A system can manipulate grief. A company can exploit attachment. Generated speech can be falsely attributed to someone who never said it. Consent can be violated. A simulation can be used to pressure the living.

These are real problems because they identify mechanisms: manipulation, commercial exploitation, false attribution, violated consent, psychological dependence, confusion between artifact and person. But the mere dynamism of preserved knowledge does not yet explain the condemnation.

People already preserve objective knowledge of loved ones. People already reconstruct abstractions of the dead. The responsive simulation makes that existing process more visible.

So the question becomes sharper:

> **What exactly has been polluted?**

If the photograph is acceptable, why does responsiveness change the moral category? If the diary is sacred, why does querying the archive become profane? If you already imagine what the deceased would say, why does externalizing part of that process become intolerable? There may be good answers. They have to identify what changes. *Creepy* captures an observation. It does not finish the explanation.

The same distinction prevents an opposite mistake. Rejecting false personhood does not require rejecting useful preservation. A letter can change your life after the author is gone. A book can create a problem centuries later. A piece of music can reorganize someone who never met the composer. A recorded lecture can introduce a criticism after the speaker is gone.

An AI-mediated reconstruction can make objective traces searchable, comparable, translatable, responsive, connected to other objective knowledge. That can increase epistemic reach without resurrecting the person. The dead do not return merely because their objective knowledge becomes interactive. But the living may be able to encounter more of what the dead left behind.

The interface changes. The ontology need not.

8. Legacy Without Transfer

Mortality often turns you toward legacy. Children, grandchildren, a family name, a company, a work, a monument, a reputation, a contribution. The language often implies continuation.

> My children are my future. My blood continues. My values will live through them.

Individuality does not permit that compression without criticism. The child is not you extended through time. The child is another person, freely and epistemically. They may inherit objective knowledge, biological knowledge, resources, problems, stories, property, institutions. They do not inherit your epistemic mind. The transfer is impossible in the strong sense. The child's future belongs to the child.

Legacy therefore has to mean something else. Objective knowledge provides the candidate. You can create something that survives your biological continuity: a criticism or theorem, a poem or private note, a design or program. A memory made explicit can travel too, as can a strange conjecture, a failed explanation carefully recorded, or a question nobody else thought worth asking.

Once externalized, the artifact can enter another epistemic mind. It can be reconstructed, misunderstood, corrected, combined, ignored, recovered later, translated, computed, fed into another technology. Its creator loses control over its reach. That is precisely what makes objective knowledge powerful.

Suppose you fear for your child's future. The temptation is to imagine control: enough money, the right school, the right values, the right institutions, the right inheritance, protection from the wrong influences. Some of these may solve real problems. But the child's future problem space cannot be possessed. Future knowledge cannot be predicted in its content. Future dangers cannot all be known. Future explanations cannot be supplied in advance.

The strongest legacy may therefore be less direct. Perhaps an idea you externalize becomes part of objective knowledge. Perhaps nobody notices. Perhaps an artificial system later ingests it. Perhaps another person encounters the idea inside a completely different problem and connects it to another explanation. The combination may eventually contribute to a technology, institution, criticism, or medical insight that changes conditions affecting people you care about.

No one may know the genealogy. The causal path may be impossible to reconstruct. This is not a prediction. It is a statement about the open-ended reach of objective knowledge. The effect does not need to preserve your name, resemble your intention, or remain inside your family. Your child cannot inherit your mind. But they can inhabit a world partly transformed by knowledge you helped make objective.

Now widen the scale. Billions of people means billions of evolved epistemic histories: different distributions of inexplicit knowledge, skills, memories, perceptual structures, problems, mistakes, unrealized conjectures, explanations waiting to become explicit.

The number itself is almost useless. You cannot apprehend billions, let alone billions of individual lives. Population abstraction returns because your access is finite. But the present limit on that access is not obviously a physical law. Your ancestors encountered fewer objective traces from fewer people. You can encounter vastly more. Future people may encounter more still.

Artificial intelligence may increase the practical depth with which one person can access objective knowledge created by many others. Not because AI merges minds. It may do the opposite. A letter hidden in an archive can become searchable. A forgotten essay can become relevant to a new problem. A language barrier can become smaller. An obscure criticism can be connected to a distant domain. A person's scattered artifacts can become a coherent surface for inquiry.

The possibility is already present. Whether we choose to build toward it is another question. Mortality makes the underlying asymmetry harder to ignore. Inexplicit knowledge can be rich, effective, beautiful, powerful. It can guide a life. It is also vulnerable because it is physically instantiated in a person. When the person dies, much of it may disappear.

Explicit objective knowledge has another possibility. It can migrate.

> Brain to speech. Speech to writing. Writing to book. Book to database. Database to model. Model to another person. Person to code. Code to institution. Institution to criticism. Criticism to new explanation.

The substrate changes. The knowledge changes too. No transmission is perfect. Every reconstruction can introduce error. But objective knowledge can travel beyond the body that first made it explicit. That gives externalization a practical significance mortality makes impossible to ignore. Not every thought should be preserved. Not every draft is good. Not every idea deserves attention. The point is not indiscriminate accumulation.

The point is uncertainty about future relevance. You cannot know in advance which strange criticism, private observation, failed solution, aesthetic experiment, technical trick, moral discomfort, or obscure question will later become useful inside another problem.

There can be personal value simply in getting an idea out of your inexplicit mind and onto an objective surface. Here we add mortality, not as an ought but an is. Sometimes the surface is the only thing that remains.

> **The tragedy is not only that a person dies. It is that knowledge can die without ever becoming available to anyone else.**

9. What Survives

Knowledge is physically instantiated, so no particular copy is indestructible. Books burn. Servers fail. Languages disappear. Archives are lost. Civilizations collapse. Memories decay. There is no need to call knowledge eternal.

Its remarkable property is different. Knowledge can migrate between substrates. It can be copied, reconstructed, translated, compressed, decompressed, re-embodied. An

explanation can outlive every person who first carried it. It can disappear for centuries and be recovered. It can enter domains its creator never imagined. It can become part of the inexplicit knowledge of someone who never learns where it came from.

This gives objective knowledge an unusual durability. You are finite. The reach of what you create need not be. That is not personal immortality.

The person you love is not valuable because they are a perfect instance of humanity, because you possess them, because they continue you, or even because you are trying to understand them. You never fully do.

Love can instead be understood, in part, as sustained engagement with an epistemic mind whose depth exceeds your reconstruction and always will.

Nothing here explains love. Biology matters. Sexuality matters. Attachment, culture, memory, beauty, physical attraction, history, objective knowledge, inexplicit knowledge, the beloved—all remain part of the problem.

Nothing here explains madness. The relationship between epistemic conflict and mental suffering remains largely unknown.

Nothing here explains death away. A grave remains a grave. No epistemology converts grief into an intellectual exercise. No argument restores the next sentence a dead person would have spoken. No theory makes the loss less real by placing it inside a causal structure.

That was never the purpose. The purpose was to deny these phenomena exemption from the same critical treatment as everything else.

Here we pushed the epistemology somewhere less comfortable: into love, confusion, grief, death. The result is not that these things have become simple. The opposite.

Love becomes stranger when the beloved is understood as an evolving epistemic mind that cannot be possessed. Humanism becomes deeper when curiosity about people follows from what people are rather than from inherited moral wisdom. Madness becomes a problem about the stability and movement of epistemic minds without becoming reducible to simple explanations. Death becomes more terrible when a person is understood not as one biological unit among billions but as an irreplaceable evolved trajectory containing inexplicit knowledge and unlimited, now unrealizable epistemic possibility. And legacy changes.

That may be the most durable contribution available to a finite epistemic person. Not continuation of the self. Not control over the future. Not possession of descendants.

Objective knowledge. An idea placed into the world. A criticism made available. A problem articulated clearly enough that another person can inherit it. A strange connection preserved long enough to meet a future mind.

Billions of people are alive inside that openness now, under widely varying constraints. Each carries biological knowledge older than civilization and objective knowledge

accumulated across millennia. Each carries an internal landscape no one else can fully reconstruct, and knowledge that has not yet become objective at all.

Death can close that path. But objective knowledge can continue beyond the person who created it. A problem made visible can enter another epistemic mind, encounter explanations its creator never knew, and become part of something neither person could have imagined.

That possibility does not begin after death. It is already present whenever you encounter the knowledge left by others. You inherit words, tools, explanations, materials, and unfinished problems. You criticize them, rearrange them, discover new relations, and sometimes bring something into the world that did not exist before.

The components may be familiar. The explanation may be new.

That is where the next problem begins.

How can what already exists become an invention that did not?

XXVI. I, Inventor



You contain more knowledge than you can say, write, or preserve. Much of it remains inexplicit. It exists in perception, memory, skill, salience, expectation, interest, taste, unfinished problems, half-formed distinctions, a sense that something is wrong, a sense that something could be otherwise.

You can carry such structures for years without making them external. Then sometimes something crosses the boundary: a note, a sketch, a sentence, a tool, a proof, a melody, a new category. Or a machine, a joke, a criticism, a way of arranging familiar things that did not exist before. The object may be trivial. It may transform civilization. Scale is not yet the important distinction. Something that had not previously existed in objective knowledge now does.

We have a familiar word for one class of such events: invention.

The word is usually reserved for conspicuous artifacts such as machines, devices, processes, patents, and new technologies. That use is legitimate, but it may be too narrow for our purpose. If you can create explanations, and explanations can become objective knowledge, then invention may describe something more general than the creation of machines. It may name one characteristic passage through which you change the objective world.

Invention gives the mystery of creativity a more concrete surface. It lets us watch some of the movement between inexplicit knowledge, explicit structure, tools, criticism, and

objective knowledge. Once that movement becomes visible, another problem sharpens. The problem is not only how new solutions appear. It is how new problems appear, how evaluative criteria acquire local meaning, and how criticism becomes creative. It is also how you can work with structures that are mostly familiar and still bring a new abstraction into the world.

1. Invention Is Not Inventory

Imagine the stereotypical inventor in the garage. There is a workbench, a notebook, loose components, a motor, wire, fasteners, a sensor, a half-disassembled machine, perhaps some software. Nothing in the room looks metaphysically mysterious. Most of the parts already exist. The materials are familiar. The scientific principles are familiar. The tools are familiar. The inventor did not create electromagnetism, steel, arithmetic, the programming language, the screwdriver, perhaps any individual component.

Yet something can happen in that room that did not exist before. The inventor arranges familiar structures under a new abstraction. The resulting object behaves differently because the relations among its components are different. A causal possibility becomes actual.

The novelty may not reside in any one part. It may reside in the architecture: the sequence, the control relation, the use of one thing as a constraint on another. A familiar component may acquire a new role. A feature previously treated as incidental may become central.

The creation is therefore not well described as the appearance of new matter. The matter was already there. Nor is it merely the accumulation of more information. The information needed to describe every component may also have been available. The new thing can be an abstraction: a new explanatory organization of existing parts.

Compare it.

The naive equation fails:

$$A + B + C + D \Rightarrow I.$$

Possession of the components does not entail the invention. If it did, every garage containing the same parts would contain the same invention. The important event is not that the parts came together. It is that a new way for them to belong together became objective.

Now remove the garage. A child draws something. The drawing may not be new to art history. It may reproduce conventions the child has absorbed without knowing their history. It may resemble thousands of earlier drawings. No patent office would care. No museum would care. No market need care.

Yet something real may have happened in the child's epistemic mind. A distinction became visible. An imagined creature acquired form. A spatial relation that had existed only vaguely became explicit. A story became drawable. A hand learned how to make a line correspond to an intention. A problem became solvable that had not previously been solvable for that child.

The drawing is objective knowledge in a modest sense. It can persist after the thought passes. It can be inspected, criticized, revisited, changed, shown to another person. It can teach the child something about what the child meant.

The child may look at the page and say:

> No. That is not right.

The page has become a criticizable surface. Enough of an internal structure has become objective that the structure can answer back.

Calling the drawing an invention may sound inflated if we hear the word only through patent law or industrial technology. But the epistemic structure is recognizable. Something not previously available to the child has become available through creation. The scale differs. The institutional status differs. The reach differs. The kind of event may not.

This gives us a first distinction:

$$\text{epistemic invention} \neq \text{institutionally recognized invention}.$$

The institutional category may identify a special class of inventions. It does not necessarily define the more general phenomenon.

The same distinction turns back onto you. You never begin from nothing either. Your language, concepts, tools, memories, skills, and inherited explanations all pre-exist the next thing you create. They constrain what can happen, but they do not specify the next abstraction.

Suppose you struggle with a problem for days and then suddenly think:

> I have it.

If the explanation is only creativity, the event has been classified without being explained.

The apparent leap may be only the visible surface of a longer transformation in your largely inexplicit epistemic mind. A memory becomes relevant. An analogy changes. A distinction sharpens. A contradiction accumulates pressure. An evaluative criterion shifts. Something peripheral becomes central.

You observe the result, not necessarily the path.

Creativity therefore remains a known unknown. Invention does not solve it. What invention can provide is a history: the notebook, the sketch, the discarded version,

the prototype, the failed test. A changed claim, a revised diagram, an observation that forced a redesign, or a criticism from another person can extend that history.

Abstract it.

That history lets us replace one dramatic picture:

$$\text{nothing} \rightarrow \text{finished invention}$$

with something more realistic:

$$O_1 \rightarrow C_1 \rightarrow O_2 \rightarrow C_2 \rightarrow O_3 \rightarrow \cdots$$

where each O is an objective partial structure and each C is some criticism, constraint, or failure.

The unknown has not disappeared. It has become more local. We can ask what changed between the partial structures, why this criticism mattered, why this feature became salient, why the problem was reformulated, why the next sketch took that form.

Invention gives creativity a longer observable shadow.

2. Getting It Out Changes What There Is to Know

The usual story of invention begins too late. It begins when you have an idea. Then the idea becomes a design. Then the design becomes an artifact. But what is an idea in that sequence? Where did its boundaries come from? Why did that problem become visible? Why did these facts become relevant together? Why did you see a relationship that had not previously been available to you?

Before the first explicit sketch appears, you already contain a vast epistemic mind. You can know that something is wrong before you can say what is wrong. You can perform a skill without explaining its control structure. You can recognize a face without possessing a proposition that performs the recognition. You can feel that an argument has failed before identifying the contradiction. You can be drawn toward one problem and not another without possessing an explicit account of why.

Abstract it.

A better starting point is therefore not:

$$\text{idea} \rightarrow \text{artifact}.$$

It is something more like:

$$K_i^{\text{inexplicit}} \rightarrow P_i \rightarrow O_1 \rightarrow C_1 \rightarrow O_2.$$

Your inexplicit knowledge shapes the problem field. A problem becomes sufficiently explicit to work on. You externalize a partial structure. That structure encounters

criticism. Your epistemic mind changes. Another structure becomes possible. The invention develops across the boundary between you and objective knowledge.

This changes the meaning of the epistemic imperative. **Get it out** can sound as if your mind already contains a finished object and the task is merely to transfer it, like copying a file. But much human expression does not work that way. You begin explaining and discover what you think while speaking. You begin writing and discover that the apparent argument depends upon a premise you do not accept. You begin coding and discover that the conceptual model lacks a required state. You begin building and discover that the elegant design is mechanically unstable.

Externalization can be transformative because the structure that becomes objective may not have existed fully even in explicit form before you attempted to express it.

Compare it:

$$\text{explication} \neq \text{transcription}.$$

Sometimes the attempt to explicate is itself part of invention. The notebook does not merely receive the idea; it participates in the causal environment through which the idea changes. The prototype does not merely instantiate the invention; it may reveal what the invention actually is. The explanation does not merely report the knowledge; explaining can create the problem that makes the next knowledge possible.

Get it out can therefore mean: externalize enough of the problem that reality can begin to answer back.

Tools matter because they change what can become objective and therefore what can be criticized. Writing preserves language beyond working memory. Drawing preserves spatial relations. Mathematics permits abstractions to be manipulated with precision. A physical prototype forces a design to encounter matter. Code forces an explanation to encounter execution. A simulation lets explicit assumptions generate consequences. A conversation makes a private conjecture vulnerable to another mind.

The important relation is not merely that a tool extends execution. A tool alters the surface on which criticism can occur.

You may be unable to hold ten candidate designs in active memory; a notebook can preserve them. You may not be able to calculate the behavior of a complex system by inspection; a simulator can expose consequences of explicit assumptions. You may forget an earlier explanation; a document can return it later.

Compare it:

$$\text{tool capability} \neq \text{explanation of creative origin}.$$

A tool can make criticism persistent, expose contradictions, enlarge search, preserve history, and return an objective structure to you in a form you did not anticipate. None

of that requires the tool to become you. It changes the reachable epistemic environment, and the changed environment can change the invention.

That makes invention recursive. You notice a problem and sketch something. The sketch reveals another problem. A crude prototype makes the problem physical. The prototype fails and exposes an assumption. The assumption becomes explicit. You change the abstraction. A second prototype works but creates a new constraint. A familiar component becomes relevant in a new way. A third design removes half the original structure. What had seemed central turns out to have been incidental.

The invention becomes clearer by repeatedly becoming objective.

Abstract it.

We can represent the relation as:

$$M_0 \to O_1 \to EE_1 \to M_1 \to O_2 \to EE_2 \to M_2 \to \cdots$$

where M_n is your changing epistemic mind and O_n is an objective partial structure.

This is not a model of brain machinery. There is no inner engineer, internal examiner, or miniature workshop hidden behind consciousness. The notation expresses an epistemological relation: you and the objective world repeatedly constrain one another, and the invention develops across that relation.

3. What Counts as an Invention?

The broader account creates an immediate problem. If every objective expression of new knowledge counts as invention, the word becomes useless. A copied sentence would qualify. So would a transcription. Every unprecedented arrangement of words would count merely because no identical arrangement had existed before.

Novelty alone is too weak. The new structure must belong to a problem situation and do more than differ accidentally. It must change what becomes possible to explain, criticize, or do.

So a provisional construction is:

> **An invention is the creation and objectification of a new abstraction that changes the available problem situation.**

The abstraction can be embodied physically, computationally, mathematically, linguistically, institutionally, or artistically. The invention need not be patentable, commercially important, or historically unprecedented. The child can still invent within their epistemic mind.

It excludes mere copying and accidental novelty without epistemic consequence. The important novelty is explanatory: a new relationship, a new constraint, a new causal organization, a new distinction, a new way of making some transformation possible.

It also clarifies why inherited knowledge does not dissolve invention. The garage contains inherited tools because civilization contains inherited objective knowledge: language, mathematics, scientific explanations, engineering standards, manufacturing techniques, software libraries, materials, design patterns, recorded failures. You create inside a world already dense with knowledge.

This can make invention look paradoxical. If the components are known, how can the invention be new?

The paradox appears only if knowledge is treated as an inventory of parts. Objective knowledge also contains relationships, transformations, constraints, and explanations. An invention can therefore be built almost entirely from familiar elements while embodying an abstraction that did not previously exist.

That is why recombination is not yet an explanation of creativity. Everything physical can be described as a recombination of prior matter. The relevant question is not whether familiar things were recombined. It is why these elements became related under this abstraction, why you saw the problem this way, why one familiar component acquired a new role.

Calling the result a recombination does not explain the abstraction that made the recombination intelligible. Nor is the invention identical to one physical token. The object may be destroyed while the invention survives in drawings and descriptions. Another person may reproduce it. The same abstraction may be instantiated in different materials. A software version may preserve the relevant structure. A later inventor may generalize it.

The invention is closer to an explanatory abstraction that can be embodied. It specifies relationships that make a transformation possible. The new knowledge resides partly in the structure of those relations.

Often the result is modest: a better hinge, a new fastener, a strange toy, a revised workflow, a trick in code. Market size does not determine the epistemic structure of the event. An abstraction has become objective that was not previously available in that problem situation.

And it may not happen in one moment. The romantic picture wants a spark, a date, a flash in which the inventor sees the whole structure. In practice, the initial idea may contain the core abstraction but fail mechanically. The first prototype may work for a reason the inventor misunderstands. A later failure may reveal the real mechanism. A critic may force a distinction that becomes central. A patent claim may isolate an abstraction that was only implicit in the drawings. A commercial deployment may reveal a second invention inside the first.

The invention has a history because knowledge has a history; it evolves. A new abstraction can become increasingly explicit through repeated cycles of objectification and criticism. The process may contain many inventive acts. Some survive. Some

disappear. Some were never as important as they felt at the time. The final artifact is the surviving trace of that trajectory.

Once objective, the invention also escapes the problem situation that produced it. Another person can reconstruct it, misunderstand it, improve it, reject it, combine it with something else, or use it for a problem the inventor never imagined. A local tool may become a platform. A convenient notation may become a way to formulate a theory. A criticism may outlive the argument that provoked it.

The creator does not control that reach. Your inexplicit knowledge does not travel directly. The objective structure does. Another person reconstructs it through another epistemic mind, and the invention can change again.

4. Patentability Makes Dynamic Evaluation Visible

Societies encounter a narrower coordination problem. Some inventions matter economically. Some require large investments. Some are easy to copy after disclosure. Some are difficult to distinguish from what already existed. Some appear new only because the evaluator lacks relevant knowledge. Some are trivial variations.

If a society wishes to assign legal rights around inventions, it needs public criteria. Patent law provides one institutional answer. The exact doctrines vary across jurisdictions. For the present epistemic problem, three familiar classes are enough: novelty, non-obviousness, and utility.

These criteria do not define invention in the broad epistemic sense. The child's drawing remains outside their intended institutional scope. A private conceptual invention may never enter a patent system. A scientific explanation may transform the world without being patentable.

The institutional problem is narrower: decide which objective technical structures should receive a special legal status under relatively stable public constraints. That stability matters. A criterion such as novelty can be written down and taught: is the claimed invention already present in the relevant prior knowledge? Non-obviousness asks whether the claimed invention is more than an obvious continuation of what was already known. Utility asks whether the invention has an operative use. Without relatively stable criterion classes, similarly situated applicants could not expect similar treatment.

But stable criterion classes do not produce static evaluation. Every application creates a new problem situation. What exactly is the invention? Which abstraction is claimed? Which prior knowledge is relevant? Does a previous artifact contain the same structure? Would a skilled person have had a reason to combine familiar elements in the claimed way? Is the claimed result merely asserted, or causally supported?

The criteria do not apply themselves. They have to become specific, and becoming specific requires explanation.

Compare it:

$$\text{static criterion class} \neq \text{static evaluation}.$$

Patent prosecution makes the difference unusually visible. The applicant presents objective knowledge: a specification, drawings, claims, perhaps examples. The applicant says, in effect:

> this abstraction satisfies the institutional criteria for invention.

The patent examiner cannot settle the matter by replying only no. A rejection has to identify a reason; it must be explained. It may show where the alleged novelty is already disclosed, explain why a previous invention bears on obviousness, question the claimed utility, or show why the specification does not support the requested scope.

The examiner therefore creates a criticism. That criticism is not the original invention, but it is still creative epistemic work because it can identify a problem that was not explicit before the examination.

The applicant then receives a different problem. Perhaps the cited reference has been misunderstood. Perhaps the claim must be amended. Perhaps the apparent invention was framed at the wrong abstraction level. Perhaps one supposedly central feature was never inventive while another, previously treated as incidental, carries the real novelty. Then the examiner receives a new problem.

The process is not merely adversarial procedure. It is a recursive exchange of creative criticism over an objective surface, the patent application.

Abstract it.

A static picture of evaluation might look like:

$$TS \to C \to \text{score}.$$

The tentative solution is fixed. The criterion is fixed. The result is classification. That picture works for some problems. If a component must fit within a specified tolerance, a measuring device may settle the question.

Open explanatory problems are different. A criticism can reveal that the wrong thing was measured or that the criterion was being applied at the wrong level. It can expose a hidden dependency, isolate an abstraction no one had noticed, or show that an apparent success solves a different problem.

Abstract it again.

The dynamic picture is closer to:

$$P_1 \to TS_1 \to EE_1 \to P_2 \to TS_2 \to EE_2.$$

The evaluative criteria constrain the cycle, but their local application must itself be constructed.

Suppose the applicant begins with:

$$P_1 \ : \text{How can this invention be protected?}$$

The examiner supplies prior art. The problem may then become:

$$P_2 \ : \text{What is actually inventive here given this prior structure?}$$

The applicant responds, and a further problem can appear:

$$P_3 \ : \text{Does the prior art really teach the claimed relationship?}$$

A later exchange may expose the deeper question:

$$P_4 \ : \text{What exactly is the operative causal structure of the invention?}$$

Novelty is still novelty. Non-obviousness is still non-obviousness. The criterion classes remain recognizable. What changes is the problem to which they are being applied. A criterion therefore does more than score a fixed object. Its application can create a new problem situation that changes what object should be evaluated next.

This is the useful content of dynamic evaluative criteria. Criteria constrain creativity, but creativity may also be required to determine how the criteria apply. The standard need not be abandoned for the evaluation to change. A criticism can force a distinction. The distinction changes the evaluation. The evaluation changes the problem.

Any system that evaluates open-ended creative work faces the same general tension. Consistency requires stable constraints. Novelty prevents the valid future outputs from being exhaustively enumerated in advance. If every acceptable invention could already be fully specified, the institution would already contain the relevant knowledge.

The workable structure is therefore not a catalog of all acceptable future inventions. It is closer to:

$$\text{stable public constraints} + \text{creative interpretation} + \text{criticizable explanation}.$$

The explanation remains contestable. The next actor may expose an error. That is how the process can remain constrained while facing an open future.

5. Criticism Can Create the Next Problem

The applicant-examiner exchange makes a broader relation hard to ignore: criticism is creative. The examiner creates a reason. The applicant creates a response. Either can

expose a problem the other did not formulate, change the abstraction under discussion, or force a reconstruction of objective knowledge. The same relation can appear between inventor and prototype, engineer and simulation, reader and text. The common structure is not opposition. It is that criticism can produce a successor problem.

Creativity therefore cannot be confined to the generation of tentative solutions. The person who says "no, because…" may create more knowledge than the person who supplied the original proposal. A devastating question can contribute the crucial invention to an explanatory trajectory because it creates the problem that makes a new abstraction necessary.

Patent practice also exposes a different distinction. Not everyone who contributes to producing an invention is necessarily an inventor of the claimed invention. A complex artifact may require technicians, programmers, fabricators, researchers, designers, managers, writers, lawyers, and testers. Someone can make a practical contribution without having created the abstraction that is claimed as inventive.

The legal details vary. The epistemic distinction is more general.

Compare it:

$$\text{enabling contribution} \neq \text{inventive contribution}.$$

This matters because technologies can displace enabling contributions without thereby explaining the inventive contribution. A machine can calculate, render, search, fabricate, optimize, translate, write, and simulate. It can remove enormous amounts of human labor from the production of an artifact. That fact alone does not answer where the new abstraction came from. The distinction does not prove that only humans can ever be inventors.

It blocks a narrower inference: causal participation in production is not by itself an explanation of inventorship.

Surface performance does not by itself reveal the machinery that produced it. Output does not by itself reveal the person. Participation in execution does not by itself reveal explanatory origin. What makes these distinctions usable is an objective surface.

The invention cannot be examined as a private feeling. It must be expressed sufficiently for another person to reconstruct it: a drawing, a description, a claim, a causal sequence, a working example. The surface will always be incomplete relative to the inventor's epistemic mind, but it can still be sufficient for criticism. A scientific paper, proof, diagram, musical score, program, machine, story, model, or patent can all serve this function. Each makes something objective enough that another epistemic mind can work on it.

The surface is not the mind. The reconstruction is not the inventor's internal state. Yet objective knowledge makes criticism possible across the gap. Private epistemic worlds remain solitary. Invention can still travel.

This also means the evaluator is inside the problem. The examiner reconstructs the invention. The applicant reconstructs the rejection. Each carries different knowledge, histories, expectations, and prior abstractions. The shared objective objects — statute, claim, specification, drawings, prior art, written rejection, response — do not produce identical minds. They do something more modest yet important: they make disagreement criticizable.

The institution does not require telepathy. It requires objective surfaces over which different epistemic worlds can collide under shared constraints. The table can be shared. The explanation of how the table applies still has to be created.

6. Invention and the Open Future

Now return to the garage and give an artificial system everything in it: every component, every relevant patent, every scientific paper, every technical standard. Add every previous prototype, known failure, candidate material, and design pattern.

Has the invention therefore been specified? No. The knowledge landscape constrains what can be created. It does not uniquely determine what will be created.

You do not create outside history. You inherit objective knowledge. But inheritance is not destiny. The same library can enter many minds. The same components can enter many workshops. The same problem can generate different explanations. The same criticism can reorganize your epistemic mind and leave another person's unchanged. The existence of the primitives therefore does not entail the new abstraction.

This is what non-obviousness tries to capture institutionally — not metaphysically, not perfectly, but recognizably. The difference between "everything was already there" and "therefore the invention was already there" is the difference between ingredients and explanation.

Of the familiar patentability criteria, non-obviousness makes that difference especially visible. Novelty asks whether the structure was already there. Utility asks whether the structure can do something. Non-obviousness asks about the path from existing knowledge to the new structure: could the claimed invention have been reached as a routine continuation of what was already known?

The legal standard is not a theory of creativity. But the problem it confronts is epistemically recognizable. If future knowledge were fully derivable from present knowledge by a fixed procedure, the future invention would already be implicit in a mechanical, executable sense. The epistemology explains instead that new knowledge leaves the future open. An explanation that does not yet exist cannot be fully consulted in advance.

Invention gives that openness a practical form. The ingredients can exist. The physical possibility can exist. The relevant person can exist. And still the abstraction does not exist until it is created.

Utility reveals a complementary relation. Useful for what? A structure becomes useful relative to a problem. Knowledge turns physical possibilities into resources by making transformations possible. The same relation appears here. A device is useful because someone can place it inside a problem situation. A mathematical technique becomes useful when a class of problems becomes visible. An artifact can remain useless until another invention changes the surrounding problem space.

Utility is therefore relational:

$$\text{structure} + \text{problem situation} \rightarrow \text{use}.$$

Invention is not a property sitting entirely inside an object. It participates in a world of problems. Change the problems and its meaning can change. An old invention can acquire new utility. A new invention can reveal a problem no one had previously formulated. The arrow runs both ways.

Novelty depends on a different relation: objective memory. You may independently create something that already exists. For you, the process may have been genuinely inventive. For the institution, the result may not be novel. Both descriptions can be true because epistemic invention and institutional patentability answer different questions.

As more prior knowledge becomes searchable, rediscovery becomes easier to detect. That can prevent some wasted effort, but the collision can also improve the problem. A prior solution may reveal that your contribution is narrower, broader, or located at a different abstraction level. The previous inventor may have solved a neighboring problem under a different description.

Objective memory does not make invention unnecessary. It changes the problem situation in which invention occurs.

7. Artificial Intelligence Enlarges the Criticizable Surface

Artificial intelligence changes the scale of that objective memory. For most of history, the practical landscape of relevant objective knowledge was sharply bounded. No inventor could read every book. No examiner could inspect every technical document. No scientist could compare every neighboring model. Every evaluator worked through compression: indexes, libraries, expertise, classification systems, search strategies, professional memory.

Now much larger regions of objective knowledge can be brought into one interaction. Not perfectly, neutrally, or without selection, but at a new scale. You can ask whether the same structure has appeared under another vocabulary or which familiar components

have been arranged differently. You can ask what criticism another field might make, which prior explanation destroys a claim of novelty, or what assumption the design is smuggling in. An evaluator can ask symmetrical questions.

Technology can therefore make parts of the evaluative surface computationally active across an immense landscape of objective knowledge.

Suppose your inexplicit knowledge has produced a weakly formed conjecture. You may discover almost immediately that the idea already existed under different terminology. That collision can destroy one problem formulation and expose another. Your real contribution may be narrower, broader, or located at an interface between two bodies of knowledge. The criticism may even reveal that there was no invention at all. Those outcomes differ, but each can improve the problem.

The expansion can also produce noise: spurious similarities, bad analogies, confident errors, premature closure, a flood of candidate criticisms that makes the important criticism harder to see. The resource is therefore not manifest truth. It is a larger criticizable surface.

That distinction becomes especially important if we imagine automating patent evaluation. Search an immense patent corpus. Retrieve similar claims. Map technical terms. Compare diagrams. Identify components. Trace support from claims to specifications. Find candidate prior art. Generate possible combinations. Preserve contradictions and return them later. These are substantial capabilities. A biological person cannot inspect the whole landscape. A computational system can traverse far more of it.

But a machine containing the criteria has not thereby replicated the process. A machine searching prior art has not thereby replicated the process. A machine generating candidate objections has not thereby replicated the process.

To reproduce the dynamic process in the strong sense, the system would have to participate in the creation of the next problem. It might have to recognize that the abstraction under evaluation had been framed incorrectly and create a new interpretation of a familiar structure. Its criticism would then have to change the problem situation, provoke a reformulation of the claimed invention, or reveal that the previous evaluative criterion had been instantiated at the wrong level.

Then recurse. That is a deeper requirement than retrieval.

A machine can carry astonishing amounts of the table: criteria, precedents, examples, technical corpora, prior arguments, rejected claims, successful amendments, language patterns, diagrams, distinctions made objective by thousands of earlier participants. That capability is epistemically important because the machine becomes a new kind of objective surface. It can return more of the inherited landscape to the current problem.

Carrying the table is still not the same as recreating the dynamic process. The stronger claim would require the machine not merely to generate text labeled criticism or retrieve

an analogous rejection, but to create the explanatory distinction that changes what the actors believe they are evaluating.

Suppose a system nevertheless produces excellent patent-examination outputs. Its searches are strong. Its rejections are accurate. Its explanations resemble those of skilled examiners. Perhaps applicants prefer them. Perhaps appeals decline. That would establish powerful technical competence. It would not establish that creativity had been explained. Successful behavior does not by itself settle the causal explanation of the behavior. A machine may participate productively in creative criticism without giving us a theory of creative criticism. Engineering can outrun philosophy.

The same caution becomes clearer if the hierarchy is kept visible.

1. External epistemology: problems, tentative solutions, criticism, successor problems. This gives the structure of knowledge growth.

2. Your internal epistemic mind: vast, path-dependent, historically constructed, partly inaccessible to introspection, capable of creating explanations. This is where the unresolved transition becomes intimate.

3. Tools: language, notes, sketches, prototypes, mathematics, code, machines, artificial intelligence. They enlarge the surfaces through which internal structures can become objective and criticizable.

4. Objective knowledge, which subsumes the tools. The invention leaves you. It persists, travels, collides, enters institutions and other minds, and generates criticism you could not anticipate.

The levels interact, but they are not interchangeable. The tool is not you. The objective artifact is not your inexplicit epistemic mind. The epistemic cycle is not a brain mechanism. The institutional criterion is not creativity.

Artificial intelligence can expand many practical stages once a problem has become available to interaction. It can help you articulate, generate variants, search, compare, simulate, criticize, remember, reconstruct earlier states, expose contradictions, translate between vocabularies, return old objective knowledge to a new problem.

It can make the loop faster, denser, less forgetful, more adversarial, more expansive — and noisier. You may confront more candidate explanations than you can evaluate, more criticisms than you can inhabit, more prior art than you can reconstruct, more variants than you can meaningfully compare.

The problem can shift from scarcity of objective knowledge to abundance without salience. That shift does not collapse the hierarchy. It makes the hierarchy more important. The tool does not own the arrows. It enlarges the surface over which you can move without, by that fact alone, explaining the transition.

8. The Depth of the Arrows

Return to the smallest marks in the familiar epistemic cycle:

$$P_1 \to TS \to EE \to P_2.$$

The arrows looked simple because the schematic is compressed. They are in reality strange and deep. Invention makes one class of them concrete.

Consider an inventor who has just received a criticism. The criticism is objective. The inventor can read it, perhaps even agree with every sentence. Still, the next move is not determined. The invention may be abandoned, narrowed, generalized, given a different mechanism, defended against the criticism, redirected toward another use, or transformed into an entirely different invention. Nothing in the arrow symbol explains which transition occurs.

The institutional process does not remove that uncertainty. It gives the uncertainty an explicit surface. The problem becomes harder to hide behind words like creativity, intuition, judgment, or insight. Candidate generation is not the deepest difficulty; it can be made cheap. Candidate testing is not the deepest difficulty either; many tests can be automated.

The deeper problem is that criticism can create the next problem, and the next problem changes what counts as a relevant solution. The state space is therefore not fixed. The evaluator is not merely moving through a known landscape. Evaluation can alter the landscape.

Where does the problem come from? Why this arrow, and why now?

Invention gives us an external history in which that mystery becomes concrete if not explained. A prototype fails. Why does this failure become the important one? A prior patent is discovered. Why does it force a new abstraction rather than abandonment? A criticism is offered. Why does it reorganize one epistemic mind while another criticism leaves the structure untouched? A familiar component is noticed. Why does it suddenly solve a problem that had seemed unrelated? The records may show what happened. They do not explain why the next problem became alive.

For you, this changes the form of the mystery. You do not experience your epistemic mind as a search through a fully specified landscape. Problems become salient. Criticisms wound one explanation and leave another intact. An old memory becomes relevant. A failed attempt makes a distinction visible. An objective artifact returns something to you that you did not know you had put into it.

The causal surfaces can often be reconstructed after the fact. Reconstruction of the surfaces is not yet an explanation of the transition.

The sharper question is no longer merely:

> How does creativity happen?

It is:

> Why does this epistemic mind produce this new abstraction through this problem now?

That question remains open. Invention has not dissolved the problem of problems. It has made more of the problem criticizable.

9. The World Answers Back

To be a person is not to be constantly inventive. It is to be the kind of entity for whom invention is possible. You are always inventive in the broader sense that you are always creative.

And invention does not merely solve problems. It can create them. A prototype reveals a hidden constraint. A new machine creates a safety problem. A new abstraction exposes a contradiction in an old theory. A criticism reveals that the inventor was solving the wrong problem. A successful invention creates new classes of problems because new possibilities now exist. This recursion is not incidental. It is part of the evolving structure of knowledge.

The patent case makes one part especially concrete. Applicant and examiner can begin with shared criteria and still have to create the actual evaluative problem together. The criterion class is stable. The problem is not. Evaluative criteria are created. Dynamic evaluation is possible because the next problem does not sit waiting as a labeled item in a fixed list. It has to become visible through criticism, failure, a tool, another person, or an objective artifact behaving in an unexpected way.

Invention changes the meaning of the epistemic imperative to get it out. The route across the boundary is not transcription. You may discover the structure while trying to express it. A tool can preserve enough of the trajectory for an earlier state to return. Objective knowledge can answer back.

The familiar garage makes the structure visible because the inventory is not enough. Patent evaluation makes the criticism visible because the criteria are not enough. Artificial intelligence can enlarge the objective surface because search and variation are not enough. At each scale, more of the trajectory becomes inspectable without the creative act itself being explained.

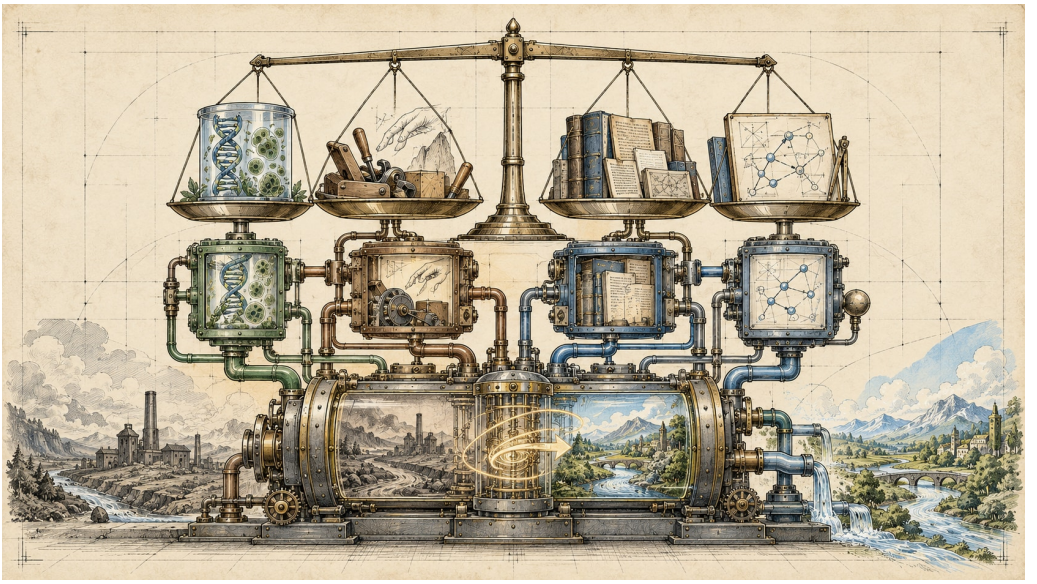
That is why invention is useful here. It prevents creativity from disappearing into a single word. An invention evolves. The history has objective surfaces. The surfaces expose failures. Failures generate criticisms. Criticisms can generate problems. Problems reorganize the path. Somewhere through that recursive movement, a new abstraction becomes objective knowledge.

The invention may depend on thousands of inherited explanations, tools, materials, and institutions. The inventor may add one distinction, one new relation, one abstraction. Without the inherited knowledge, the invention would not exist. Without the new abstraction, it would not exist either.

The contribution can be small in visible form while changing what the entire inherited structure makes possible.

What would it mean to say that the inherited knowledge did most of the work, while the new explanation contributed only a little? How do we weigh knowledge?

XXVII. What Knowledge Powers



Suppose we take everything explored so far seriously. Genes contain biological knowledge. You inherit a body shaped by evolution. Much of what you know remains inexplicit. You enter a world saturated with objective knowledge: language, norms, institutions, technologies, scientific explanations, moral explanations, stories, categories, laws, and interfaces. And you can create explanatory knowledge that did not previously exist.

You can concede all of this and still reach a radically different conclusion about what you are. You may still feel mediocre. You need only change the weights.

Perhaps biology does most of the work. Perhaps culture does. Perhaps inexplicit knowledge does. Perhaps explanatory creativity is real but comparatively minor, a thin surface floating on a much larger inherited structure. Perhaps the explanatory you has been given a superordinate place only because this inquiry happens to value explanation. Perhaps you are mostly genes and society, with creativity contributing the last few percentage points.

Nothing in that objection requires denying any of the categories. It accepts them, then rearranges them. That makes the objection more serious, because it exposes another abstraction we have not yet criticized.

What could those allocations possibly mean?

1. More of What?

The language arrives easily. Biology matters more. Culture matters less. Social conditions overwhelm individual creativity. Explanatory knowledge is important, but only at the margin. Most of what you know was not created by you. Most of what you do depends on knowledge you cannot explain. Most people do not create scientific theories, new technologies, political institutions, or philosophical systems. So perhaps explanatory creativity is real but small.

The intuition is powerful because ordinary comparison often works this way. One object is heavier than another. One road is longer. One population is larger. One source explains more variance in a statistical model. One expense occupies a larger share of a budget. In each case, an abstraction supplies a common dimension and makes comparison possible.

But before asking whether biology or explanatory creativity contributes more to you, a prior question appears:

More of what?

Mass has kilograms. Distance has meters. Variance has a defined statistical structure. Money supplies a common economic abstraction for some otherwise heterogeneous things. What common unit could assign shares of you to a genome, a language, a motor skill, a social norm, a scientific explanation, a memory, and a newly created moral criticism? No such unit has been supplied.

Perhaps one can be created. But creating it would itself require an explanation. We cannot smuggle the unit into the question and then treat the answer as measurement.

A population statistic can be real. A variance component can be real. A trait score can be useful. None of that establishes that you are internally composed in the same proportions as the representation. The decomposition belongs first to the model. Reality does not owe the model matching seams.

2. What Knowledge Does

Part of the confusion comes from grammar. We speak of knowledge as though it were something accumulated in a container. You have knowledge. A civilization stores knowledge. A genome contains knowledge. A book preserves knowledge. A skill embodies knowledge.

The noun is useful, but it can mislead. Knowledge is not a substance poured into different reservoirs. It has no mass, and there is no balance on which biological knowledge can be placed on one side and explanatory knowledge on the other. Yet knowledge is not merely an abstraction invented for convenience. It does something. That is why the concept matters.

Knowledge is roughly **physically-instantiated information with causal properties**. The phrase can be made less mysterious by separating its parts.

Information means that some physical structure carries a difference. Different nucleotide sequences are different. Different sentences are different. Different neural states are different. Different arrangements of a tool are different.

Causality means that the difference can participate in producing another difference. A genetic sequence participates in constructing and regulating an organism. A learned skill changes what you can do. A traffic rule changes how people coordinate movement. An explanation of infection changes which actions a surgeon takes. A mathematical notation changes which relationships become tractable.

The information matters because it participates in change. That causal property distinguishes knowledge here from information in the broadest sense. A random arrangement may contain information. It does not follow that the arrangement solves a problem, guides a transformation, or participates in a knowledge-bearing structure.

Knowledge does causal work. That gives us a better question than how much it weighs:

> **What changed because this knowledge was present?**

The answer need not be enormous. Knowledge often works incrementally. Before a distinction exists, two things are treated as one; afterward, they can be criticized separately. Before you learn a skill, an action cannot be performed reliably; afterward, it can. Before a notation is invented, a relationship may be difficult to manipulate; afterward, it becomes available to thought. Before you understand that an inherited rule can be questioned, the rule may appear as part of reality; afterward, it can become a problem.

These are increments, but **increment** is not a universal unit or an atom of knowledge. It names a local analytical cut in a problem situation: something was possible before, and something else is possible now.

The change may be small while its consequences are not. This is another appearance of the **Fallacy of Mediocrity**. A tiny physical intervention can redirect an enormous system. A few lines of code can alter what millions of people encounter. A short proof can reorganize a mathematical field. A new distinction can make an inherited moral category unstable.

Visible amount does not determine explanatory significance. Neither do age, frequency, the number of people already carrying the knowledge, or the size of the institution in which it is embedded.

The useful question is local:

> **What work is this knowledge doing in this problem?**

That question does not solve the problem. It gives criticism somewhere to begin.

3. Knowledge Changes Form

Biological knowledge is old. Some of it is unimaginably old. The structures involved in metabolism, perception, movement, reproduction, immune response, and neural development did not appear with language or science. They are products of evolutionary histories extending far beyond your life.

Objective human knowledge in the epistemic sense is younger. Writing is younger still. Formal science younger again. Modern computation younger again. A new idea may be seconds old.

It is tempting to turn chronology into hierarchy. Ancient knowledge appears deeper, more fundamental, more causally important. Sometimes it is. But age is not a measure of epistemic significance, and causation does not run only from ancient biology toward newer culture. People remain evolving animals, while the knowledge people create changes the environment in which later biological evolution occurs.

Fire makes the interaction vivid. Humans inherited biological digestive systems. At some point, humans also acquired and retained knowledge of controlling fire. Cooking became objective knowledge embodied in practices, tools, social transmission, hearths, and food preparation. Part of food processing moved outside the body, changing the environment in which human digestion operated.

Over sufficiently long periods, objective knowledge created by people participated in changing the conditions under which later biological knowledge was selected. The change required no foresight, and we do not possess a complete account of every mechanism involved.

The important relation is simpler: the categories interact. Biological knowledge participates in producing organisms capable of creating objective knowledge. Objective knowledge changes the environment in which biological knowledge is selected. The causal history loops.

Language exhibits the same general pattern. Biology makes language possible. Language becomes objective knowledge. Language changes which ideas can be expressed, preserved, criticized, and transmitted. Those changes alter institutions, social coordination, reproductive patterns, survival, education, technology, and environments. The new environment becomes part of the conditions under which later biological and cultural variation occurs. There is no clean boundary where one source finishes and another begins.

The interaction also happens inside your life. You learn something explicitly and practice it until the knowledge changes form. The musician no longer needs to state every movement. The driver stops calculating every steering adjustment. The speaker stops consciously reconstructing every grammatical rule. The explicit can become inexplicit.

The reverse movement happens too. Something feels wrong. A skill fails. An expectation is violated. You become uncomfortable with a norm you once enacted automatically. Something that was operating inexplicitly becomes available for explication.

But explicitation is not the discovery of a hidden sentence waiting fully formed inside your mind. You create an explicit representation of what may have been guiding you. That representation can be wrong. It can be criticized. The criticism can then alter future inexplicit knowledge.

Explicit knowledge can become inexplicit; inexplicit knowledge can become a problem for explicit thought; objective knowledge can reorganize both; biological and social structures continue to participate throughout. The forms remain distinguishable without becoming independent.

It is easy to draw these transformations too neatly:

Biology produces a person. The person creates an explanation. The explanation becomes objective knowledge. Objective knowledge changes society. Society changes later persons.

The arrows make the story intelligible. They can also create an illusion. To write

$$A \to B$$

is not to explain how A relates to B.

Why does one observation become interesting? Why does one tension become a problem? Why does one problem produce this conjecture rather than another? Why does a criticism that was available yesterday reorganize your mind today? Why can a creative leap feel discontinuous even when many explicit and inexplicit dimensions have been changing incrementally?

The arrow names a change. It does not necessarily explain the machinery of the change. That ignorance matters. If we cannot completely explain many of the transitions among forms of knowledge, assigning precise relative causal weights becomes stranger still. We would be quantifying a decomposition whose internal transformations remain partly unexplained.

This does not weaken the causal problem. It expands it. We know something changed. We know knowledge participated. We can sometimes isolate enough of the change to criticize an explanation of it. But much of the machinery remains unresolved.

The swamp remains swampy.

4. The Swamp and the Model

The metaphor of solid epistemic ground is attractive. You discover facts, place them on a foundation, add more, and build upward. Some claims receive more support than

others. Some sources become stronger. Eventually the structure stands. That is not our epistemic reality.

Your epistemic mind is shifting. Biological dispositions coexist with social norms. Explicit explanations coexist with inexplicit expectations. Old categories remain active after you have criticized them. You may explicitly reject tribalism and still experience tribal intuitions. You may understand the asymmetry of criticism and still search for confirmation. You may know that a trait category compresses individual causal histories and still use the category automatically. You may value freedom and still become uncomfortable when another person's freedom generates outcomes you dislike.

The old knowledge has not vanished, and the new knowledge has not simply replaced it. They can coexist, conflict, recombine, become more or less salient, reach temporary equilibrium, and disturb it again.

The **swamp** names this movement. A perfectly static epistemic landscape would be more troubling. No contradiction would become a problem. No variation would survive long enough to become a conjecture. No old expectation would collide with a new explanation. No criticism would reorganize anything.

Knowledge evolves because structures can vary and because some variations can be eliminated. Externally, this occurs across people, institutions, theories, technologies, and generations. Internally, it occurs across explicit and inexplicit structures inside you. The swamp is not the bane of knowledge. It is one of the conditions of knowledge becoming.

Earlier we distinguished reality from the epistemic mind through which you encounter it. Call reality R. Call your evolving epistemic mind M_i. The distinction is easy to state:

$$M_i \neq R.$$

Living that distinction is harder. You do not ordinarily experience a model hovering between you and the world. You do not look at a stranger and experience, *my current compressed representation of this person, constructed through biological, cultural, perceptual, linguistic, and inexplicit knowledge, suggests danger.* You experience danger. You do not encounter a social rule and experience, *an inherited abstraction is currently organizing my expectations.* You experience what people are supposed to do. You do not experience gravity as a theory. You experience the cup falling.

Your epistemic mind feels like reality because it is the interface through which reality becomes available to you. That interface includes language, memory, skills, expectations, social categories, scientific explanations, moral explanations, inexplicit reactions: everything that makes the world intelligible enough to act within it.

The compression is necessary. Without it, finite cognition would drown in detail. But the same necessity creates a danger: a new explanation can fit your existing compression so well that the fit itself feels like truth.

5. The Truck

Imagine that you feel territorial. The feeling may concern property, family, a neighborhood, a nation, a social group, or a relationship. Perhaps it is strong. Perhaps it has always seemed obvious.

Then an explanation arrives: humans evolved territorial and coalitional tendencies because those tendencies solved recurring problems of survival and reproduction in ancestral environments. The explanation may come with measurements, cross-cultural patterns, behavioral experiments, genetic associations, neural correlates, and evolutionary models.

You feel a sudden compression. Of course I feel this way. My ancestors evolved these tendencies. Science has measured them.

The explanation lands inside an epistemic mind already prepared to receive it. The feeling was there. The social categories were there. The common sense was there. The language was there. Now a scientific explanation appears to bind them together. A truck has just driven through your epistemic mind.

Nothing developed so far automatically stops it. You may let the truck through, or you may stop it and inspect the cargo. The difference need not be intelligence, education, or skepticism as a personality trait. It may be your problem situation.

If your problem is *Why does this feeling exist?*, the evolutionary explanation may suffice. But does its origin make the feeling a good moral guide? Does a population regularity explain this particular person? What was measured, and could different mechanisms produce the same result? Does an evolved disposition justify a legal rule? Now the truck is open.

You cannot inspect every truck. A finite person depends on compression. You board a train without reconstructing its engineering, speak without tracing every word's history, and cross a bridge without recalculating its loads. Life depends on knowledge remaining compressed until a problem gives you reason to inspect it.

So the implication cannot be *inspect everything*. The problem is to identify when the compression itself has become part of what you are trying to understand.

If the evolutionary explanation helps explain a passing feeling and nothing important turns on it, perhaps no further decomposition is needed. If the explanation is being used to classify a child, write a law, condemn a group, design an institution, excuse violence, define a gender, or predict a person's limits, the problem has changed. Now the contents of the truck matter.

The same objective knowledge can therefore require different degrees of inspection in different problem situations. The problem situation changes what needs to be criticized. It does not change reality.

6. Better Explanations Without a Scale

If we cannot weigh different forms of knowledge, and if the significance of a piece of knowledge depends partly on the problem situation, are we left with arbitrary preference?

No. Two explanations can address overlapping problems and still differ radically in quality.

Evolution and creationism both speak to the history and diversity of living things. That overlap does not make them epistemic equals. One invokes mechanisms whose consequences constrain what can happen: variation, selection, inheritance, retention. Those mechanisms can collide with observation and with other explanations. The other can vary its central cause to accommodate almost any possible outcome. Whatever happened can be said to have been intended, allowing the explanation to absorb criticism rather than be changed by it.

The problem situation is shared. The explanations are not equally good by epistemic standards.

Comparison therefore does not require a numerical scale of truth. It requires criticism. One explanation may solve problems that another merely renames. One may require arbitrary changes whenever criticism appears. One may contradict other explanations we have better reason to retain or create new problems with greater reach.

A useful word for the direction of this improvement is **verisimilitude**. The term need not carry more substance than the present problem requires. We cannot measure how much truth an explanation contains. We can still move toward better explanations by discovering and eliminating errors.

The same caution applies to **reach**. A good explanation developed for one problem can constrain many others. Evolution reaches beyond one species. Computation reaches beyond one machine. An explanation of abstraction can change how we think about psychology, law, technology, morality, and personhood.

The reach can be enormous while the physical expression is tiny. A sentence can have more reach than a library of unrelated descriptions. That does not mean the sentence contains more kilograms of knowledge. Reach tells us something about the structure of an explanation; it does not convert heterogeneous knowledge into shares of you.

The Fallacy of Mediocrity appears again. Visible volume is not explanatory significance. Frequency is not explanatory significance. Historical duration is not explanatory significance.

A forest and a match make the asymmetry vivid. The forest may have grown over millennia. The match is tiny. Yet the match can burn the forest.

The image has limits. Knowledge is not combustible matter. An explanation does not contain more energy because it has greater epistemic reach. Common sense is not literally destroyed when a criticism appears. The relation worth preserving is narrower:

Causal consequence need not track apparent quantity.

Suppose people possess some evolved tendency toward forming in-groups and distrusting outsiders. The exact biology is not the point. The tendency may exist inexplicitly, participate in common sense, become encoded in language, be reinforced by norms, formalized by institutions, preserved in laws, and celebrated in stories. A child may therefore inherit biological knowledge and enter an objective environment that appears to say the same thing.

Then someone creates an explicit criticism:

> Perhaps the distinction between our kind and the others is not a reliable guide to moral standing.

The criticism is physically small: perhaps a sentence, a conversation, a book, a legal argument, or a protest sign. Its size tells us almost nothing about what it can do. It can make a previously invisible structure available to criticism. Other minds can reconstruct it. Some reject it. Some change. Institutions and language may change. The social environment of later children may change. Over enough time, even the biological selection environment may change.

The match did not outweigh the forest. It introduced a transformation. Nor does the forest disappear. Old knowledge remains active. You may endorse equality and still feel tribal pull. An institution may abolish one discriminatory rule and preserve another. Old common sense can regrow. The metaphor works only if the burning remains incomplete: knowledge does not become clean. It becomes different.

7. What Explanatory Knowledge Changes

The deeper version of that transformation sits inside the epistemology itself. Evolution creates knowledge through variation and selection. Variants arise. Some disappear. Some persist. What persists changes the conditions under which later variation occurs. Long before explanatory people existed, biological knowledge grew this way.

Epistemology inherits that evolutionary structure and changes something within it. You can conjecture, then deliberately seek error. Selection can retain what survives. Criticism can ask what is wrong with what survived. The orientation changes. A successful explanation is not protected because it succeeded. Its success can instead expose it to harder criticism.

This is not psychologically automatic. We like confirmation. We search for support. We feel reassurance when an explanation fits. Even if you reject justificationism, you can continue seeking justification inexplicitly. The older tendency can persist beneath the newer explanation.

But the explicit distinction changes what becomes possible. You can notice the justificatory impulse as a problem. That is a small increment with potentially large consequences.

This is where an artificial symmetry among forms of knowledge would mislead.

Inexplicit knowledge can learn, guide action, and solve problems. Biological knowledge can embody solutions no organism understands. Social knowledge can coordinate millions of people. But while a structure remains inexplicit, it can be difficult to criticize *as knowledge*. You may feel friction, failure, discomfort, or conflict without yet having an explicit object that can be inspected.

Explicit explanatory knowledge changes that architecture. Once enough of a structure becomes explicit, it can be stated, compared, criticized, recombined, taught, externalized, preserved, passed to another mind, and returned later in altered form.

The difference is not epistemic mass. It is access to deliberate error correction. An impulse can become a question: *Why do I feel this?* A norm can become a problem: *What problem was this rule solving?* The same operation can reopen a scientific claim: *What does this measurement actually establish?* It can also reopen a self-description: *Is this category explaining me or merely compressing observations about people like me?* Explanation creates a surface on which error can work.

Science enlarges that surface deliberately. Its power does not lie simply in accumulating more observations, in special authority, or in an imaginary method that converts data into truth. It lies partly in the systematic creation of explicit explanations that can be made vulnerable to criticism.

Measurements, experiments, statistics, and models matter inside problem situations shaped by explanations. An unexplained regularity remains a regularity. A precise measurement remains a measurement. The scientific achievement becomes deeper when the structure that generated the observations is made explicit enough to criticize.

Science therefore expands the surface of error correction. It takes parts of reality that previously appeared as common sense, mystery, habit, or inexplicit expectation and turns them into problems. The process compounds: one explanation makes another possible, one instrument creates observations unavailable before, one notation makes new relationships visible, and one criticism reorganizes a field. Progress accumulates not as weight but as transformed possibility.

Sometimes bodies of knowledge developed through different problem trajectories begin to fit together. An explanation in biology becomes compatible with one in computation. An explanation of the person stops conflicting with physics. A moral explanation

stops requiring a theory of mind incompatible with psychology. An internal conflict between experience and explicit explanation dissolves because a better abstraction reorganizes both.

We call this **consilience**. The plain-language relation comes first: several structures that were developed separately can become mutually compatible. The name compresses that temporary fit.

Consilience is therefore not merely internal consistency, accumulated evidence, or consensus. It is a provisional equilibration. The equilibrium matters because contradiction has temporarily diminished; the provisionality matters because a new criticism can make one piece impossible and move the landscape again.

Consilience belongs naturally inside the swamp. It is not solid ground. It is temporary structure within movement.

8. When Knowledge Structures Meet

The same process occurs inside you. You are not one coherent model. Many structures can be active in the same epistemic mind: a scientific explanation, an inherited norm, an inexplicit fear, a moral commitment, a practiced skill, a childhood category, a new abstraction, a remembered failure.

Some never meet. Some reinforce one another. Some produce conflict. Some become explicit only because another structure collides with them. The pattern is not a list of independent contents but an **internal ecology**: interacting knowledge structures whose effects depend partly on what else is already present.

That is why the same explanation can enter two minds differently. You may hear an evolutionary explanation of territoriality and experience consilience because it fits feeling, language, common sense, and prior scientific knowledge. Someone else may experience tension because the same explanation collides with a moral problem, a causal distinction, or a prior criticism.

The explanation has not changed. The problem situation has. No single external description of the explanation therefore tells us what epistemic work it will do inside you. Reconstruction is local, and the consequences are path-dependent.

You cannot leave your epistemic mind. You cannot step outside every abstraction and compare with reality. You cannot inspect every piece of inexplicit knowledge, reconstruct every causal pathway, or eliminate compression.

Yet you can sometimes make the compression itself a problem. *Of course* can become *Why does this feel obvious? Everyone knows that* can become *What exactly do we know?* The same move can reach more elaborate compressions. *Science has shown this* can become *What was observed, what was modeled, and what is being explained? This is just*

human nature can become a request to separate biological knowledge, social inheritance, interpretation, and any new abstraction built from all three.

You may answer incorrectly. The criticism may fail. The old explanation may survive. But the truck has stopped. Its contents are now available to criticism. That is not omniscience. It is **epistemic autonomy**: not freedom from inherited knowledge, but the possibility that inherited or compressed knowledge can itself become an object of criticism.

The problem situation supplies the selection pressure for criticism. What are you trying to understand? What are you trying to change? What contradiction has become active? What harm might follow if the compression is wrong? What new possibility appears if the abstraction can be reopened? A choice of sandwich does not require reconstructing nutritional science. A dispute over whether an inherited racial category should determine legal rights presents a different problem. A driver can rely on an interface that an engineer investigating a failure may need to reopen. A patient may reasonably use a medical category that a researcher, asking whether it compresses several different causal structures, may need to decompose.

Criticism is not a ritual. It is work directed at problems. Context changes what requires criticism without changing reality itself. Once a conjecture is available, its origin can become another part of the problem. Sometimes it matters whether an idea arose from biological intuition, social pressure, a scientific paper, an AI system, a dream, a mathematical derivation, or a memory. Origins can expose constraints, inherited assumptions, conflicts of interest, or hidden structures.

But origin does not determine epistemic standing. A biological intuition is not true because it evolved. A social norm is not true because millions of people inherited it. A scientific explanation is not true because a scientist created it. An AI-generated conjecture is not false because a machine produced it. A dream is not true because it felt profound.

Once the conjecture exists, another question becomes central:

What do we do with it?

Can it be made explicit? What problem does it address? What causal work does it claim to describe? What criticisms can reach it? What survives? What changes?

Its ancestry remains available when relevant. But criticism is not genealogy.

9. Knowledge Becomes

Now return to the noun. Knowledge persists. A genome persists across generations. A skill persists across repeated action. A book persists on a shelf. A law persists in an

institution. An explanation persists in objective form. Without retention, there is no cumulative evolution.

But retention is only part of the story. Knowledge also varies, combines, becomes explicit, becomes inexplicit, changes environments, creates new problem situations, changes which observations matter, creates new abstractions, generates contradictions, reaches provisional equilibrium, and breaks equilibrium.

The noun hides the motion. Knowledge is not merely something that **is**. Knowledge **becomes**. The grammar is imperfect. The claim is not that knowledge ceases to be objective structure. It is that the epistemically important structure is evolutionary: retained knowledge becomes the starting point for further change.

You are therefore not a container holding different shares of knowledge. You are one place where knowledge can become. That changes the moral problem explored through this series. Knowledge is treated conjecturally because it may contain error. Its current success does not entitle it to permanent protection. If persons are knowledge-creating entities, then treating a person as finished would freeze the very process the explanation attributes to personhood.

Your present explanation does not exhaust your future. A type does not contain every explanation you may later create. Your current knowledge need not remain your final knowledge. The relevant moral conditions are therefore those under which you can encounter criticism, create alternatives, reject explanations of you, and become different.

The moral explanation does not come first. It follows from the account of the person as a fallible creator of explanatory knowledge: an open-ended problem world whose future knowledge cannot be completely derived from its present state.

The resulting principle can be compressed as:

> **Preserve the conditions under which the knowledge-creating process can continue.**

The epistemology refuses to freeze an explanation. Applied to a knowledge-creating person, it supplies a reason not to freeze the person either.

Compression now returns in a sharper form. A personality score, diagnosis, demographic category, legal status, résumé, reputation, or self-conception can all solve problems. They can coordinate, predict, simplify decisions, and make institutions possible. The error begins when the compression is allowed to become the person.

Here we add another version. Even this ontology of knowledge can become a compression: biological knowledge, inexplicit knowledge, objective knowledge, explanatory knowledge. The categories are useful because they make distinctions visible. If we turn them into weighted ingredients of personhood, however, the abstraction hardens and reproduces the original error in more sophisticated language.

The categories were created to help explanation. They are not justified.

10. A Different Basis for Comparison

So what remains? We cannot weigh the forms of knowledge. We cannot separate them. We do not fully understand many of the causal transitions among them. We cannot inspect all of them. We cannot escape compression.

That can sound like epistemic defeat. It is not. The failure of a common scale does not remove comparison; it changes what comparison must track.

Take whichever knowledge structure is relevant to the present problem. It might be a biological disposition, an inexplicit expectation, a social norm, a scientific theory, a moral criticism, a mathematical abstraction, or a technical design. Then ask:

> **What work is this doing?**

What changed because it was present? What problem does it solve? What problem does it create? What does it make possible? What does it make invisible? What happens when the explanation of its role is criticized? The unit is not a chunk of knowledge. It is the **increment of causal work** that can be made sufficiently explicit to criticize.

Now return to the original objection: perhaps explanatory knowledge contributes very little compared with biology or society. What problem is that comparison solving?

If the problem is prediction of variance in a defined population measurement, a statistical decomposition may be exactly the right tool. If the problem is explaining why you experience territoriality, the same decomposition may be inadequate. If the problem is explaining why territoriality became a moral problem for you, another explanation is required. If the problem is explaining how a criticism spread through institutions and changed later environments, another level appears. If the problem is explaining why the criticism was created at all, we reach the creativity gap again.

The answer changes because the problem changes. Reality does not become relative. The explanatory task becomes more precise.

11. The Weight of Explanatory Knowledge

We can now assert something that initially seemed forbidden: explanatory knowledge is more powerful. That claim does not mean it outweighs biological knowledge or constitutes a larger percentage of you. Nor does it mean explanatory knowledge always has greater causal effect or that an explicit idea will defeat an inexplicit disposition. Its power lies elsewhere.

Explanatory knowledge can expose knowledge to error correction. It can make an inherited structure criticizable and create a problem where previously there was only

common sense. It can cross minds as objective knowledge, preserve criticism beyond one life, and change the criteria by which later explanations are evaluated.

Because those changes can be retained and criticized again, they can compound. You learn an explanation and criticize it. The criticism is externalized. Another person inherits the problem and creates another explanation. A field, institution, or technology changes. The environment changes. Later persons begin somewhere else.

The power is recursive. That is why explanatory knowledge occupies a special place in this ontology, and why you do too. Biological knowledge does not need to know that it has a problem. A norm can persist without explaining itself. An inexplicit skill can work without becoming known. An institution can preserve a rule long after the problem that created it has disappeared.

You can make such inherited knowledge into problems. You can ask why a reaction feels obvious or what problem a rule was solving. You can ask whether that problem still exists, what an abstraction hides, or what would have to be true for an explanation to fail.

You are not outside biology, society, objective knowledge, or causality. You are physically embodied in all of them. But you are one place where knowledge becomes an object of explanation and criticism. The inherited can become problematic, the obvious strange, the compressed decompressible, and the next problem visible. In that sense, you are one place where knowledge can become a problem for itself.

That is not quantitative superiority. It is an exceptional epistemic function.

12. Preserve the Movement

The impossible question—*How much does this knowledge weigh?*—can be replaced by a more useful one:

> **What changed because this knowledge was present in this problem situation?**

The change is often incremental even when the explicit experience is not. A new explanation can suddenly reorganize years of inexplicit development. A tiny criticism can expose an error distributed through common sense, language, norms, and institutions. The size of the criticism does not measure the size of the consequence. The Fallacy of Mediocrity applies to knowledge too.

Inside you, the problem becomes harder because your epistemic mind does not feel like a model. It feels like reality. A compelling explanation can align with biological feeling, social expectation, personal history, objective science, and existing abstraction all at once. The fit can produce the most dangerous phrase in epistemology: *of course.*

Nor can you inspect everything. A problem situation makes some compressions worth reopening and leaves most others in place. When inspection begins, no numerical scale of truth is required. Explanations can be criticized for arbitrariness, contradiction, missing mechanism, inflated scope, or failure to solve the problem they appear to address. Error can be found and eliminated. That is enough verisimilitude for the present problem.

Explanatory knowledge is special because it can make knowledge itself criticizable. It can turn the obvious into a problem, externalize criticism, cross minds, and compound. Science enlarges this surface deliberately. Objective knowledge preserves it. Tools extend it. Persons create within it.

The way knowledge is treated and the way people are treated are not separate stories once people are understood as the exceptional creators of explanatory knowledge.

The corresponding moral implication is not to mistake a current compression for the whole person or a present explanation for a final identity. The conditions under which you can criticize, reject, revise, create, and become otherwise are conditions under which the knowledge-creating process can continue.

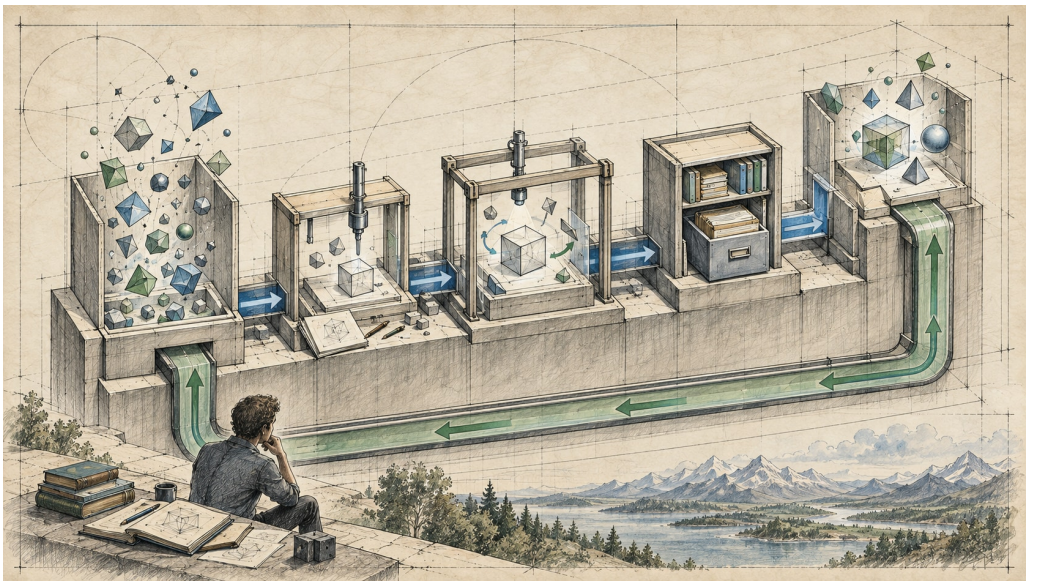
Perhaps the deeper error was not that we tried to weigh things. Weighing made knowledge look still and justified. Knowledge is not only what has become. Knowledge is becoming; you are one of the exceptional places where it becomes.

Knowledge is not only what has become. Knowledge is becoming, and you are one of the places where it becomes.

But if knowledge changes what other knowledge can do, then its development cannot be represented merely by adding another item to an inventory. The relations change. The problem changes. What becomes possible changes.

What does knowledge look like while it is moving?

XXVIII. You Are a Deep Loop



Knowledge is not a substance. We cannot weigh biological knowledge against explanatory knowledge as though both occupied a common scale, or divide you into percentages of genes, culture, inexplicit knowledge, and explanation. The categories are useful because they make distinctions visible. They become dangerous when the distinctions harden into ingredients.

Not:

> how much knowledge is there?

But:

> what changed because this knowledge was present?

That question moves. Knowledge enters a problem situation and the situation is no longer the same. Something becomes possible. Something becomes impossible. A distinction appears; an old distinction disappears. A criticism becomes available. A problem becomes visible. The next state is not the previous state plus one more object. The relations have changed.

We've encountered this problem through the arrows between epistemic states of problems, tentative solutions, and criticism; through invention. A note changes you as

you write it. A prototype changes what you can see in the problem. A criticism changes the abstraction being criticized. Objective knowledge returns to you and alters what you can do next.

We called it recursive and emergent. Both are useful. Both are dangerous. Either can make an explanation sound deeper without explaining anything. So the problem here is simple to state:

What does knowledge look like while it is moving?

1. Move One Word

Consider a sentence.

> The dog bites the man.

Now move three words.

> The man bites the dog.

Nothing supernatural happened. The same three nouns and verbs remain. The letters did not acquire new physical properties. No new vocabulary was introduced. Yet the meaning changed because the structure changed.

Grammar makes the relation unusually visible. The symbols do not carry the meaning independently. Their relations matter.

Try another.

> Only she told him that I lied.

Move one word.

> She only told him that I lied.

Move it again.

> She told only him that I lied.

Again.

> She told him only that I lied.

The physical change can be tiny while the semantic change is large. A few marks move across a page, their relation changes, and the sentence now says something different. That gives us a concrete place to begin. A system can contain substantially

the same components while expressing something new because the relations among the components have changed.

It does not make relations magical. The relation is physically instantiated. A word occupies a different place. You process a different syntax. Different interpretations become available, and different consequences can follow. If the sentence is an instruction, different actions may occur. If it is a confession, different people may be blamed. If it is a law, different conduct may become prohibited.

Meaning has causal consequences because the structure carrying the meaning changed. This is ordinary, which gives us a useful constraint. When something genuinely new appears, we do not need to imagine that a new substance entered reality. Sometimes the novelty lies in the organization.

Now consider a chair.

Remove one leg. The chair is different and probably worse at being a chair. The difference is real, but difference alone is not creation in the sense that has concerned this inquiry. You changed the object. You did not necessarily create a new problem-solving structure.

Remove the back. Perhaps the remainder works as a small table. Again, structure and function have changed, but a flat surface elevated from the floor was already available as an abstraction. You have repurposed the object.

Now flip the original chair upside down. Put it in a gallery. Give it a title.

> *The Tyranny of Sitting.*

Perhaps this is terrible art. That is not presently the problem.

Something different has happened. The chair still consists of wood, fasteners, joints, surfaces, and empty space, but it has entered another structure. Its orientation and context changed. Its title changed. Its relation to the viewer, to other objects, to inherited artistic conventions, and to the ordinary function of chairs changed.

The object can now participate in a criticism. Someone may laugh. Someone may be irritated. Someone may produce an essay about furniture and power. Someone else may produce a better artwork in response. The chair did not contain these consequences as hidden material properties waiting to be released.

The new organization changed what the object could mean and do inside an epistemic mind. Nothing magical occurred. The chair moved. The gallery admitted it. Someone named it. Someone encountered it. Relations formed. The resulting structure became capable of participating in problems that the chair in the dining room did not.

Known components. Inherited tools. Existing knowledge. A new relation. The inventory does not explain the result.

But neither does the phrase:

> something emerged.

That word is often used exactly where an explanation becomes difficult. Cells interact; consciousness emerges. Individuals interact; society emerges. Neurons interact; meaning emerges. Simple rules interact; complexity emerges.

Sometimes this is harmless shorthand. Sometimes it is where explanation stops. We describe the components, describe the result, and put *emergence* between them:

$$A + B + C \to \text{emergence} \to D.$$

The mystery has not been solved. It has been labeled.

That would be especially damaging here. This inquiry has repeatedly refused to convert ignorance into ontology. We do not know how explanatory creativity works. Calling creativity emergent does not improve the situation. We do not know why a particular problem becomes salient. Calling salience emergent does not explain it. We do not know the causal path by which an inexplicit tension becomes an explicit criticism. Calling the transition emergent does not tell us what happened.

The useful part can be kept without treating the word as a cause. A new organization can appear through transformations among structures that already exist, and the new organization can possess causal or semantic properties that the components did not possess separately. We may call the appearance of that organization **emergence**.

But emergence is not cause. The transformations remain the explanatory problem.

Physical realism constrains the account. Whatever happened, happened in reality. No semantic spirit entered the sentence. No artistic essence entered the chair. No epistemic substance entered your brain. There were physical states, transformations among them, and relationships instantiated in those states.

If our explanation contains a gap, the gap remains ours. Reality did not skip the middle merely because our account did.

Incrementalism helps here. Earlier essays used it in social and technological contexts: changes can be local enough that their consequences remain criticizable and errors remain detectable. The same idea protects novelty from mystification.

Abstract it.

Suppose a new explanation exists at t_2 that did not exist at t_1. That is genuinely interesting. But it does not follow that reality jumped discontinuously between the states. There may be a path:

$$S_1 \to S_2 \to S_3 \to \cdots \to S_n.$$

Every local transformation can be physically ordinary. A memory becomes active. A sentence is reread. A distinction changes. A contradiction becomes salient. A word is written, and the written word alters the problem. Another explanation becomes relevant. A criterion changes. You sleep. The problem remains. You return. A new relation becomes explicit.

At the end, you say:

> I have an idea.

The idea may be genuinely new. Nothing requires it to have existed as a hidden complete object at the beginning, and nothing requires a magical event at the end. The problem is to explain the transformations.

Often we cannot. The arrows between epistemic states are abstractions. They represent movement we have good reason to believe occurs. They do not claim to identify the neural machinery by which one thought becomes another.

Emergence does not close the arrows. It says that the resulting organization cannot be identified merely by listing the components that preceded it. That is useful. It is not an explanation of creativity.

The same pressure falls on compression. A map preserves selected relations while discarding enormous amounts of detail. A personality trait compresses many observations into a dimension. A model makes a system tractable by omitting structure. Compression helps explain why abstractions are powerful and why they can hide what matters.

But compression is not the whole story. We did not create the new sentence meaning by merely removing information. We rearranged the structure. We did not create the artistic claim by merely subtracting pieces of chair. We placed the object in new relations. An inventor may use parts that all existed before, while the invention is not explained by the inventory.

Separate those two operations. One removes structure while preserving what matters for a problem. The other reorganizes structure so that a relation becomes visible, objective, or operational in a form that was not previously available.

They can combine. A new abstraction may discard enormous detail while exposing a relation nobody had previously made explicit. The equation can be shorter than the observations while explaining more. The criticism can collapse a thousand pages into one distinction. Compression may participate, but the novelty lies in what the new organization can now do.

That gives us another constraint:

$$\text{difference} \neq \text{novel knowledge}.$$

Randomly shuffle the words in a sentence and you will produce novelty in the trivial sense that the arrangement did not previously exist. Most arrangements will be nonsense. Break every leg off the chair and the resulting object is new in one physical sense. That does not make it an invention.

The epistemically interesting case is when a new organization changes the problem situation. A new criticism becomes possible. A new action, question, transformation, or relation becomes available. Something enters the causal world of knowledge that was not there before.

2. Evolution Is a Verb

This brings us back to evolution. The word is easy to hear as a noun: a topic in biology, a theory of speciation, a historical process that happened over immense periods before you were born. That is all true. It is also a compression.

The deeper structure is active. Variation occurs. Some variants persist and others disappear. What persists changes the starting conditions under which later variation occurs. The cycle continues. Evolution is not merely the history of biological change. It is change acting upon the conditions of later change.

Biological knowledge makes this unusually concrete. A mutation changes a structure. The structure changes the organism. The changed organism encounters the world differently. Its survival and reproduction occur under those conditions. Retained structures become part of the starting point from which later variants arise. The output returns as a condition of the next transformation.

Abstract it.

We can write a crude version:

$$V_1 \to S_1 \to R_1.$$

Variation. Selection. Retention.

But retention is not an endpoint. The retained structure becomes part of the next state:

$$R_1 + V_2 \to S_2 \to R_2.$$

Now the recursive structure is visible. The result of one pass participates in the conditions of the next.

Nothing requires foresight. The selection mechanism need not understand what it retains. An eye can become better at forming useful images without any organism possessing an explanation of optics. A wing can become better suited to flight without any organism understanding aerodynamics. Biological knowledge accumulates without explanatory criticism. That is one of evolution's great explanatory achievements.

Now turn the structure toward epistemology as an explanation of knowledge. You also vary. Conjectures appear. Actions are tried. Explanations are criticized. Some structures persist and others disappear. What persists changes what you can do next. The resemblance to biological evolution matters, but it can hide the most important difference.

Consider the biological sequence:

variation → selection → retention → changed conditions for later variation.

Now consider an epistemic sequence:

$$P_1 \to TS_1 \to EE_1 \to P_2.$$

Problem. Tentative solution. Error elimination. Successor problem. Here too, what results becomes part of what happens next. But epistemic selection can do something biological selection does not need to do: it can become a problem for itself.

You propose an explanation. Criticism exposes an error. The explanation changes. That is familiar. But the criticism can also expose an error in the criterion by which you were judging the explanation.

You can say:

> I was asking the wrong question.

Or:

> I thought this counted as evidence, but it does not.

Or:

> I was evaluating the solution by a criterion that only made sense inside the old problem.

The process has not merely selected among variants. It has altered something about selection itself.

$$EE_n \to \text{change in the conditions governing } EE_{n+1}.$$

You can criticize the criteria of criticism. You can make the problem itself problematic. You can reject an inherited evaluative standard. You can construct a new distinction that changes which variants are even visible as variants. This does not make epistemic evolution supernatural. It makes the selection mechanism different and presently mysterious.

Biological selection can accumulate knowledge without knowing that selection is occurring. Epistemic selection can produce explanations about selection. That explanation can become objective knowledge, and objective knowledge can alter later selection. The process turns upon itself.

3. The Old Problem Is Not the New Problem

Now recursion has earned its name. A recursive epistemic process is not merely a loop:

$$A \to B \to A.$$

That would be repetition. The interesting structure is:

$$A_1 \to B_1$$

followed by:

$$(A_1 + B_1) \to A_2.$$

The output returns, but because the output exists, the next input situation is different. An inventor makes a prototype. The prototype fails. The failure becomes objective knowledge. The inventor encounters that knowledge and is now different. The next prototype is created from a problem situation that includes the failure of the first.

The object changes the inventor. The changed inventor changes the object.

$$M_0 \to O_1 \to EE_1 \to M_1 \to O_2.$$

This structure makes invention concrete. The deeper point is that the output does not merely come back. It returns to a system that has already been changed by producing it. The loop changes the thing that loops.

Now let the thing that loops be you. You are stuck. A problem has occupied you for days. Perhaps you are writing, designing, trying to understand a conflict, debugging a program, or making a decision. Something does not fit. You think about it and nothing happens. So you leave. You read something unrelated. You have a conversation. You sleep. You walk. You become interested in something else. Then you return, and the problem looks different.

We often say:

> I needed time to think.

Or:

> I had to let it gestate.

Or:

> the answer came to me later.

Those phrases are useful, but they can hide the process. The problem did not necessarily sit unchanged inside a sealed compartment until an answer appeared.

You changed. Knowledge entered. Other knowledge disappeared from attention. Memories shifted in salience. Other problems became active. New relations formed. Inexplicit knowledge may have changed before any new explicit sentence appeared. The old problem then encountered a different epistemic system.

Abstract it again.

We can represent the movement crudely:

$$(P_0, K_0) \to K_1.$$

Then:

$$(P_0, K_1) \to P_1.$$

Then:

$$(P_1, K_1) \to K_2.$$

The notation is not your mind. It only prevents us from imagining repetition where change is occurring.

You say:

> I kept thinking about the same problem.

Perhaps. But the problem at the tenth encounter may no longer be the problem at the first. One assumption has become visible. One memory now matters. One distinction has disappeared. One criticism has acquired force. One criterion no longer feels compelling. The language remains the same. The epistemic object has moved.

This does not make every repeated thought productive. A fear can recur. A grievance can become more rigid. An obsession can narrow the available problem space. A circular explanation can immunize itself against criticism. Recursion is not necessarily good. Evolution is not necessarily progress. Emergence is not necessarily improvement.

The question remains:

> what is changing?

A repeated thought can return to a nearly identical structure and strengthen it, or it can re-enter a changing system and acquire a new relation. From outside, the difference may be difficult to see.

One process repeats:

$$P \to TS \to P \to TS \to P.$$

Another moves:

$$P_1 \to TS_1 \to EE_1 \to P_2 \to TS_2 \to EE_2 \to P_3.$$

Both may look stuck. Both may talk about the same subject and return to the same memory. Inside, one structure may be hardening while the other is changing. The observable category does not tell us which. The arrows can change the conditions under which later arrows appear. The landscape is landscaping itself.

That makes another familiar question look strange: what is the essence of a person? It cannot be biological, psychological, or social essence. You are not the Big Five, a diagnosis, a demographic category, a genome, a role, or a fixed collection of values.

Essentialism cannot be rescued by becoming more sophisticated. Perhaps your "essence" is massive a compression of your exceptional epistemic mind. Perhaps you are the kind of entity that instantiates:

$$P_1 \rightarrow TS \rightarrow EE \rightarrow P_2.$$

We have found a deep structure. But what have we actually explained?

Knowing the structure of a process does not tell us what happens when the process is turned on. You can instantiate the same abstract schema as another person while beginning with different objective knowledge, memories, bodies, inexplicit expectations, problems, criticisms, and criteria. The first transition changes the conditions of the second. The second changes the conditions of the third. The trajectories can diverge radically and rapidly.

A structural account may be true and still fail completely as an explanation of its content. That is not a defect in the structure. It is one of its consequences. A recursive process cannot be understood merely by freezing the process into a noun. You are not an epistemic cycle stored in a body. You are the cycle moving through an evolving and historically particular world while changing the world through which you move.

Even the compression, **universal explainer**, can become misleading if read as an essence. The phrase does not identify a hidden substance shared by people. It identifies a capacity, and a capacity in motion produces trajectories whose contents do not yet exist.

Your future explanation cannot be extracted from the definition of the explainer. If it could, it would not be new.

4. Objective Knowledge Enters the Loop

Your recursive epistemic movement does not occur alone. You live among objective knowledge: books, interfaces, stories, laws, scientific theories, images, algorithms, schools, tools, other people, artificial intelligences. These structures do not merely supply information to an otherwise unchanged you. They can alter the conditions under which you create knowledge.

Writing changed what could be remembered. Notation changed what could be calculated. Scientific instruments changed what could be observed. Software changed what could be searched and simulated. Artificial intelligence changes what can be generated, compared, criticized, preserved, and returned.

Objective knowledge becomes part of the recursive environment. An external artifact enters. You change. The changed you creates or selects another artifact. That artifact returns. The process continues.

This is why technological arguments framed only as exposure can become misleading. How many hours? How many messages? How many videos? How many prompts? How much screen time? The measurements may be useful, but they observe the outer surface of a recursively changing relationship. The same artifact can become distraction, instruction, friendship, obsession, a tool, a criticism, a new problem, or a way out.

The external object does not uniquely specify what is happening inside your epistemic mind. That does not make technological effects unreal. It makes the explanatory problem harder.

The same difficulty appears in the language of technological lock-in. A child is locked into a platform. A population is locked into an algorithm. A worker is locked into an interface. You are locked into a media environment. The image is spatial: a prison, walls, someone inside. But what exactly is locked? Behavior? Attention? A social network? A category? A preference? A memory? A habit? A problem situation? A selection criterion? A set of available conjectures?

Once those possibilities are separated, possible mechanisms become visible. A system can constrain exit. It can make alternatives expensive. It can hide available options. It can repeatedly return one class of objective knowledge while suppressing another. It can force you to represent yourself through fixed categories. It can reward one kind of expression and punish another. It can make criticism difficult to discover or costly to express.

But an active epistemic subject complicates the prison. You can notice the walls and criticize them. You can repurpose a tool, use one part of a platform to learn how another works, or build a new community inside a structure designed for another purpose. You can leave. You can fail to leave. You can want to stay. You can create a problem nobody who designed the system anticipated.

The prison may still be real, but we cannot infer it merely from the shape of the technology. We have to explain which epistemic transformations have become unavailable, costly, or invisible. That is a stranger prison. You may participate in constructing the loop that constrains you, while also being the entity within it capable of making the loop itself a problem.

That changes the moral question, if you choose it. Earlier parts derived a familiar principle: preserve the means of criticism. Freedom of speech, association, conscience, exit, institutional openness, and the ability to change your mind preserve conditions under which error correction can continue.

But the moving epistemic subject adds a complication. You cannot exercise a criticism that does not yet exist for you. You cannot leave by a path you cannot yet imagine. You

cannot reject a category that has not yet become problematic. You cannot formulate the alternative problem that has not yet become visible.

The formal permission to criticize may be present while the criticism is absent. That does not justify giving an authority power to manufacture the correct criticism on your behalf. The authority also lacks the uncreated knowledge.

The moral problem is therefore deeper: what conditions leave enough openness for new criticism to become possible without presuming that anyone can know in advance what the criticism will be?

The word *emerge* must remain disciplined here. A new criticism does not appear from nowhere. It may grow through conversation, failure, boredom, exposure to an alien explanation, an inexplicit discomfort, a mistake, a technology, or a child asking the wrong question. The conditions matter. But no inventory of the conditions contains the criticism as a predetermined conclusion.

You can become able to think something you could not think before. That is the moral significance of your moving epistemic mind.

5. When the Representation Returns

The same structure becomes technological once a system represents you and returns that representation to the environment in which you continue to act. A static model says: the user has preferences; the student has a level; the patient has a profile; the worker has a role; the customer has a segment; the child has a measured ability. The system learns the state and serves the state back.

This can work. Recommendation, personalization, classification, and stable interfaces are useful because you cannot reconstruct every problem from first principles at every moment.

But recursion introduces a different question. What happens when the representation re-enters the process that produced it?

Abstract it.

A system classifies you. The classification changes what you see. What you see changes what you do. What you do becomes new data. The new data strengthens the classification.

$$C_1 \to E_1 \to B_1 \to C_2.$$

Classification. Environment. Behavior. Updated classification.

The loop is not automatically bad. A language tutor can adapt to a learner and improve instruction. A medical device can adapt to changing physiology. A creative tool can remember a project and return relevant material.

The question is what can still change inside the loop. Can you surprise the model? Can the category be reopened? Can the problem be reformulated? Can the history be inspected? Can an earlier state return? Can the system preserve contradiction rather than averaging it away? Can you reject the representation? Can the system support a transition that the previous model of you would never have predicted?

A faithful epistemic technology cannot merely preserve what you know. It must avoid making your previous state into a cage for your next one.

Computation makes the possibility unusually concrete because representations can be stored, copied, compared, simulated, recombined, and returned. A computational system can preserve states with a fidelity biological memory cannot.

Imagine working on a problem with an external system. The system does not merely store the latest answer. It preserves:

$$P_1, TS_1, EE_1, P_2, TS_2, EE_2.$$

You can return to an abandoned conjecture or compare an old criterion with a new one. You can recover the point at which the problem changed, notice that a criticism now matters that previously did not, or ask what disappeared from the current representation. The technology can make movement objective. It does not create the movement. It gives movement a surface.

Invention reveals something similar. A sketch preserves enough of an earlier epistemic state for the changed inventor to encounter it again. The old artifact becomes new input. Computation can expand this capacity dramatically.

The danger is symmetrical. A system that preserves movement can also freeze it. It can compress you into one profile or treat the latest state as the true state. It can predict the next problem from the previous ones and quietly suppress what falls outside the prediction. It can interpret inconsistency as noise instead of potential criticism.

Then the technology does not merely represent you badly. Its representation can become causal. You begin to move through a world partially structured by the system's compressed account of who you were. The abstraction returns. The loop closes.

Now widen the loop once more. A new explanation is created. At first it exists only in one person. Then it is written and becomes objective knowledge. Someone else reads it; their problem changes. They build a tool. The tool changes what other people can do. An institution adopts the tool and changes its procedures. Children grow up inside those procedures. For them, the structure may no longer appear as an invention. It appears as the environment.

The organization that emerged at one point has become a condition of later variation. That is exactly what evolution teaches us to expect: retained knowledge alters the environment in which later knowledge develops.

Epistemic evolution adds another turn. A later person can reconstruct the history. You can inherit such a structure and ask why it is here and what problem it was solving. You can ask whether that problem still exists, what the structure made possible, what it suppressed, and what happens if it changes. The environment can become an object of explanation. The inherited can become a problem. The process recurses again.

This is why the repeated contrast between nouns and verbs matters. Person. Knowledge. Culture. Intelligence. Education. Technology. Evolution. Mind. Each noun is useful because it gives us a handle. Each can also freeze movement.

Evolution is especially revealing. Heard as a noun, it sounds like an object of knowledge: a scientific subject, a chapter in a textbook. Heard as a process, it is retained structure altering the conditions of later variation.

Knowledge does something similar. A noun suggests possession. *Knowing* still sounds like a state. But the movement is richer: problems become conjectures; conjectures become criticisms; criticisms become new problems. Inexplicit knowledge becomes explicit. Explicit knowledge becomes objective. Objective knowledge returns and changes inexplicit expectations. Biological knowledge builds a person capable of criticizing biological explanations. An institution embodies an old solution until someone makes the institution itself a problem.

The categories remain useful. Reality keeps moving across their boundaries. The movement is primary. The nouns are our handles.

6. But the Children

Have you heard? Children are getting dumber. This time we can measure it.

You've heard some version of the story. International tests, teenagers across dozens of countries, declining scores in mathematics, reading, science. Reading is especially worrying: not just recognizing words, but connecting sources, evaluating claims, thinking critically. Exactly what everyone says matters in the age of AI.

Meanwhile, students use chatbots for homework. Phones distract classrooms. Children read less, struggle to concentrate, and let machines answer before learning to remain inside a question.

The decline began before AI, but that detail fades as the story travels. Scores down. Screens everywhere. Attention collapsing. Machines doing the thinking.

The pieces fit together. The explanation feels obvious.

We built extraordinary machines. Then we put them in the hands of children. Now the measurements are arriving. And the books, and the panic, and the policies.

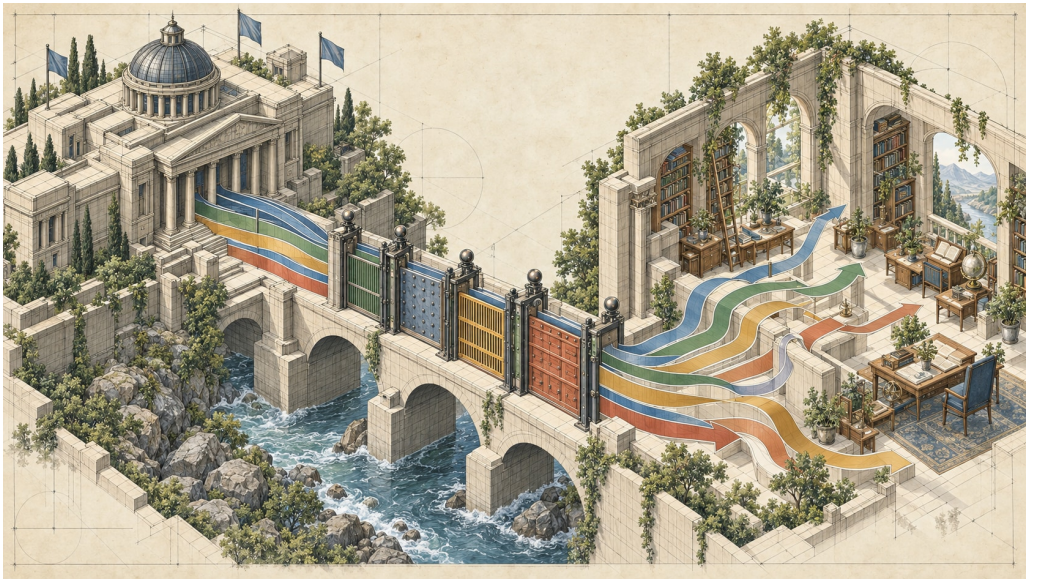
The tempting move is obvious: put the technology on one side, the test scores on the other, and call the relation an explanation. The knowledge of knowledge in motion makes that move harder. Much harder.

A changed outcome is not yet a causal path. Exposure is not yet an epistemic trajectory. A category is not yet the movement inside it. If the children have changed, what changed in the recursive conditions under which they learn, criticize, remember, attend, conjecture, and return to problems?

The truck is hurdling toward us. There is a brick on the accelerator. It is full of children.

Do you really want to stop it and inspect what's inside?

XXIX. Your Open Mind and Its Rulers



We ended with a familiar story about children, declining educational performance, digital devices, and artificial intelligence. Pieces of research, classroom observations, parental worries, and public commentary had become one explanation:

> The children are getting dumber.

The truck was coming toward us, and the final question was:

> Do you really want to stop it and inspect what's inside?

We did not. That may have seemed strange. We have spent a long time stopping trucks: a personality score, a heritability estimate, a diagnosis, a political category, a social abstraction, a value, an inherited theory of the person. Again and again, the compression became the problem, so we opened it.

Why leave this one moving? Perhaps we should have stopped it. Perhaps we still should. But not because stopping every truck is what a rational error-correcting person always does. There is no such rule.

You cannot inspect every abstraction, reopen every category, reconstruct every causal chain, or hold every problem active at once. Criticism is work. Attention is finite. Time

passes. Other problems exist. A building may be burning. Someone may need help. Your child may be calling you. The explanation you are reading may not be the problem that matters to you now.

Morality has been waiting inside this distinction. And this time the truck is ours.

1. The Missing Arrow

We have made moral claims, sometimes quietly and sometimes explicitly. Preserve criticism. Do not freeze the person. Do not mistake a population model for an individual. Do not convert an abstraction into authority. Preserve the ability to reject, revise, create, and become otherwise. Permit disagreement. Preserve exit. Avoid coercion. Treat the current explanation as provisional.

These claims often appeared to follow from the epistemology. If knowledge grows through criticism, then institutions should preserve criticism. If persons create explanatory knowledge, then persons should not be treated as finished objects. If future knowledge cannot be predicted, then your future should not be inferred from your present category. If social abstractions compress, then coercive action based upon them should be approached carefully.

The fit has become increasingly strong. Perhaps by now the moral consequences feel obvious. That is precisely when criticism becomes useful. A compression can become a truck of its own. Consilience can feel like inevitability. Reach can feel like proof. A model that reorganizes many problems can begin to feel like reality without mediation. Then the word appears:

> of course.

Of course people should be free. Of course criticism should be preserved. Of course authority should be limited. Of course coercion is wrong. Of course individuals matter.

Perhaps. But we have spent too much time criticizing manifest truth to claim exemption for its own morality. So stop this truck too. Open the compartments. What exactly follows?

A familiar philosophical warning says:

> You cannot get an ought from an is.

A description of reality does not by itself contain a moral command. A statement about what exists does not logically entail what anyone should do. That objection matters here because the epistemology has described people as fallible creators of explanatory knowledge, criticism as part of knowledge growth, coercion as capable of blocking criticism, social categories as compressions, and individual futures as open. It has then repeatedly approached claims such as:

preserve criticism; preserve individuality; preserve the ability to change.

Where did the word *should* come from? There is a tempting reply. You cannot get an ought from an is, but you cannot get an is from an is either.

Consider observations. The sky appears blue. A needle moves. A population score declines. A planet appears in one place rather than another. Someone behaves differently than expected. No finite collection of observations contains its own explanation. The explanation must be created.

$$O_1, O_2, O_3 \nRightarrow E.$$

The observations constrain the explanation. They can refute it, create problems for it, or become inexplicable under one theory and intelligible under another. But accumulation does not manufacture the theory.

Something happens between:

observation

and:

explanation.

The same problem appears between:

explanation

and:

ought.

The epistemology may severely constrain what a moral explanation can coherently say. It may expose contradictions, reveal the causal consequences of coercion, show that an abstraction does not contain the person it classifies, or explain why criticism matters. None of those statements constrain your next action.

Again, the arrow matters. We have encountered this structure before. A problem does not mechanically generate its tentative solution. A tentative solution does not mechanically generate the criticism that will matter. A criticism does not mechanically generate the next problem. A body of objective knowledge does not mechanically generate an invention. A moral explanation does not mechanically generate obedience.

We write:

$$P_1 \to TS \to EE \to P_2$$

But the arrow is not a conveyor belt. Relations change. Salience changes. A contradiction becomes visible. A new abstraction appears. Something becomes a problem that was not a problem before.

The moral transition is another instance of that structure. Suppose an explanation implies that some intervention will damage your capacity for criticism. Suppose the explanation survives criticism. Suppose you understand it. Does that force you to act?

No. You may reject the problem or rank another above it. You may disagree about the cost. You may possess knowledge I do not. You may see a conflict I have not noticed. You may be wrong. So may I.

The explanation can constrain your problem situation. It cannot occupy the place from which your next explanation will be created.

2. Finite Attention

Return to the children. The educational measurements may matter enormously. The decline may represent real harm. The role of schools may be changing. Digital media may be involved. Artificial intelligence may be involved. The tests may be compressing several transformations into one score. Some explanations may be excellent; some may be terrible.

Why not stop the truck immediately? Because the existence of an available criticism does not establish that this is the problem you should solve next. If you are an education researcher, stop the truck. If you are a teacher watching students struggle, stop the truck. If you are a parent deciding what environment to create for a child, stop the truck.

But if you are standing beside a burning building, perhaps the truck isn't your immediate problem. You have not become epistemically inactive because one abstraction remains unopened. Your attention may simply be somewhere else.

This matters because moral language often conceals the opposite assumption:

> If you understood, you would act. If you really knew, you would agree. If you cared, you would choose correctly. If you were rational, you would see what must be done.

Each statement may occasionally be true. None is entitled to be true in advance. An unopened truck does not prove indifference, irrationality, or moral failure. It tells us almost nothing about the problem currently active inside your epistemic mind.

Now imagine that the building really is burning. Smoke is coming through the hallway.

Someone says:

> We should leave.

Another person asks:

> But what caused the fire?

The question matters.

Later. Right now the causal history of ignition is not the active problem. The immediate problem is:

$$\text{people in danger} \rightarrow \text{reduce danger now.}$$

The person who pulls the alarm does not need a complete theory of combustion. The person who carries a child outside does not need to settle a philosophical argument about authority. You act under incomplete explanation constantly. That is not a defect in explanation. It is one of the conditions under which explanation exists at all.

The problem situation a causal structure that appears temporal. Some criticisms can wait; some cannot. Some actions preserve the possibility of later criticism; some destroy it. If the building burns with everyone inside, there will be fewer future explanations of the fire? No! People will have died.

So the moral objection to authority cannot simply be:

> no one may ever tell another person what to do.

A firefighter may tell you to leave. A parent may stop a child from running into traffic. A court may restrain someone who is attacking another person. An institution may temporarily close a bridge because it is failing.

The relevant question is not whether authority exists. It plainly does. The question is what work it is doing.

3. Authority as a Relation

Authority is another noun. The grammar makes it sound like a possession: a person has authority, a court has authority, a parent has authority, a government has authority, an expert has authority. Something one actor has and another lacks. But authority moves with the problem situation.

A surgeon may have relevant authority in an operating room without becoming your political ruler. A parent may decide where a four-year-old sleeps without acquiring the same authority over a forty-year-old child. A judge may impose a sentence within a legal procedure without becoming the final authority on physics. You may reasonably defer to another person's technical knowledge without transferring ownership of your epistemic authority.

Authority is better understood as a conjectural relation inside a problem situation. Who may act? On what? For how long? Under what constraints? With what avenues for criticism? With what possibility of reversal? What happens if the explanation of the authority fails?

The same institution may exercise reasonable authority in one domain and unreasonable authority in another. The same parent may correctly prevent immediate physical danger and incorrectly infer that a child's unusual interest is harmful. The same state may coordinate traffic rules and also attempt to decide which ideas citizens may consider.

The category does not decide. The explanation does, and the explanation remains fallible.

This may sound unstable, even swampy. Perhaps we want authority precisely because we do not wish to reopen every decision. That is true. A society cannot deliberate from first principles at every traffic light. A hospital cannot renegotiate sterile procedure during every operation. A child cannot independently reconstruct all electrical safety knowledge before being told not to insert metal into an outlet.

Compression is necessary. Institutions are partly machines for preserving useful compression. Authority can preserve it too.

The problem begins when temporary compression becomes ontological entitlement: a rule created to solve one problem becomes proof that the ruler knows better in general; a temporary restriction becomes a theory of the restricted person; useful expertise becomes immunity from criticism; a necessary emergency action becomes a permanent political arrangement.

Abstract it.

The transition is familiar:

$$\text{successful local solution} \rightarrow \text{generalized authority} \rightarrow \text{manifest legitimacy}.$$

That arrow is not automatic either.

Reasonable authority therefore has a peculiar structure. It can coordinate without pretending to be infallible, act without claiming final knowledge, compress without claiming the compression is the whole person, and temporarily close a decision while preserving a route by which the decision can later become a problem.

A verdict that can never be reopened is different from one that ends the present dispute while remaining embedded in institutions capable of criticism. A teacher who says, "for this exercise, use this method" is doing something different from one who says, "this is the only way your mind should work." The same distinction appears when a parent says "not now" rather than "never become someone I cannot understand."

The difference is not softness. It is epistemic structure.

4. The Removable Ruler

Political philosophy often begins with a question:

> Who should rule?

The best person? The wisest? The majority? The educated? The virtuous? The property-owning? The workers? The experts? The people? The question invites a particular kind of solution: find the correct ruler, then give that ruler sufficient authority to rule correctly.

But the epistemology explored here begins elsewhere. Every ruler is fallible. Every institution compresses. Every political explanation can contain error. Every majority can be wrong. Every expert can misunderstand the problem. Every good intention can become an environment in which criticism is expensive.

Another question then becomes available:

> How can bad rulers be removed?

The difference is larger than it looks. The first question seeks a source of correct authority; the second assumes error. The first asks how to select the right answer in advance; the second asks how to keep a wrong answer from becoming permanent. The first tends toward legitimacy as identity:

> this actor is entitled to rule.

The second tends toward legitimacy as corrigibility:

> this arrangement remains usable while its errors can be exposed and changed.

Elections are one technology for doing this. Courts can be another. Appeals, free speech, opposition parties, independent institutions, exit, competing jurisdictions, public criticism, and rules restricting arbitrary power can all create surfaces upon which bad outcomes become problems. None guarantees good outcomes. That is not their deepest function here.

Seen this way, some moral axioms associated with modernity begin to look different: human equality, freedom of conscience, tolerance, individual rights, freedom of speech, limits on arbitrary government, pluralism, consent. These are often presented as principles, sometimes as self-evident, sometimes as products of culture, sometimes as inherited commitments that modern societies simply happen to prefer. But they can also be approached as inventions.

A moral invention need not be a sentence no one had ever uttered. It can be a new distinction, a new problem, or a new relation among existing explanations. Something previously treated as ordinary becomes criticizable. A ruler's religious authority becomes a problem. Slavery becomes a problem. The legal subordination of women becomes a problem. Punishment for dissent becomes a problem. The inherited status of a child becomes a problem. The assumption that one group can determine the lives of another becomes a problem.

The new moral explanation enters objective knowledge and changes the problem environment. Institutions change. Language changes. Law changes. Children inherit different categories. Later criticism begins somewhere else. This does not prove that every liberal institution is good, that humanism is complete, or that modern societies are morally final. It explains how moral knowledge can move.

Not:

> how much does knowledge contribute?

But:

> what changed because knowledge was present?

Apply that here. What changed because freedom of conscience existed as a concept? Because legal equality existed? Because consent became a moral category? Because arbitrary detention became something that required explanation? Because a child came to be understood as a person rather than as property? Because a ruler's authority could itself become a problem?

The answer is not merely that people came to believe new propositions. New criticisms became possible. New harms became visible. Old arrangements lost their status as nature.

You could say:

> you may have power over me, but you have not explained your right to use it this way.

That sentence contains an invention. It separates:

power

from:

moral authority.

Once separated, they can be criticized separately. That is causal work.

5. Humanism Without Declaration

Humanism can be approached in the same way. One could begin with the command:

> every human life has equal dignity.

That may be an excellent moral conjecture. But we took an epistemic route.

It began with a population statistic. Then the person expanded. Genes became biological knowledge. Behavior became a surface. The trait became a compression. The mind became a vast epistemic landscape, much of it inexplicit. Objective knowledge crossed into it. You reconstructed inherited explanations, created new problems, externalized criticism, used tools, built institutions, loved, invented, changed. Another person could not fully reconstruct your epistemic mind. Neither could you. Yet you could sometimes stop the truck.

Now choose one stranger. You may know almost nothing about them. Perhaps you see them on a train, encounter them for a moment in a crowd, learn that they live on another continent, or read about someone who lived two thousand years ago. What evidence would establish that their epistemic mind is shallow? You cannot see most of it. Neither can they. The absence of access tells you little about the depth.

No command is required. Notice instead what the explanation has done to the stranger.

There is a phrase that can sound morally dangerous:

> I am not like you.

Another:

> We are not the same.

Much moral language tries to overcome this by emphasizing common humanity: shared needs, rights, vulnerability, biology, reason. There is truth in each. But the epistemology has produced a stranger conjunction. Every known person belongs to the same explanatory class. Every person is capable of creating explanatory knowledge. Yet every person's mind landscape is historically singular.

The universality is common. The individuality is not.

You and another person can share language, family, country, profession, religion, politics, childhood home, and decades of memory, yet still discover that the same event inhabits two radically different explanatory worlds. The overlap can be enormous. The divergence can be larger.

So perhaps:

> We are not the same

is not the threat. The threat appears when difference is forced into hierarchy.

$$\text{difference} \nRightarrow \text{different moral worth.}$$

That arrow must also be explained.

Earlier we described each person as a universe. Not a separate reality. There is one reality. But each person contains an enormous evolving abstraction of it.

Imagine two such universes. One contains knowledge you admire; the other, explanations you despise. One has produced great kindness; one, cruelty. One has encountered criticism and changed; one appears closed. One is brilliant in a domain you understand; one seems confused. One is a child; one is dying.

What unit would let you weigh their capacity to create the next explanation? There is none available.

The absence of such a unit does not prove moral equality. It removes an easy route to moral inequality. A demographic category cannot supply it. Neither can an intelligence score, a personality profile, a political identity, a diagnosis, or a reputation. None of these compressions contains the future epistemic trajectory of the person. The moral conclusion is still not forced. But some shortcuts have disappeared.

There are billions of people alive now. The number is easy to say; the thing it refers to is almost impossible to imagine. Billions of histories. Billions of partly inexplicit worlds. Billions of memories whose significance cannot be externally recovered. Billions of networks of salience, skill, expectation, and attachment. Fear, curiosity, resentment, humor, unfinished problems, half-formed distinctions, private contradictions. Billions of possible future explanations.

Some will never become objective. Some will become jokes, tools, criticisms, equations, acts of cruelty, acts of care, institutions, mistakes, or ideas that reorganize civilization.

The abstraction *humanity* is tiny beside what it compresses. We need it; no person can encounter everyone individually. But when the abstraction begins to feel complete, something has vanished.

Perhaps awe is the feeling produced when the compression briefly fails. Not mystery as an excuse. Not mysticism. Not reverence demanded by doctrine. The recognition that the object before you has more explanatory depth than your model can hold.

One person can sometimes do that. The class of all people is beyond imagination.

6. Judgment Without Completion

The word *evil* compresses too. That does not make it useless. Some people deliberately torture others. Some build institutions that preserve suffering. Some exploit vulnerability.

Some kill. Some terrorize. Some knowingly block criticism because criticism threatens their power. The causal consequences can be terrible.

A moral framework that cannot condemn or stop such actions would be defective. But a second claim does not follow from the first:

this action must be stopped

does not entail:

we possess the final explanation of this person.

Earlier we described evil as something like persistent harmful closure:

harmful explanation + power + blocked criticism → persistent avoidable suffering.

That remains useful. It identifies something we may need to oppose immediately without requiring an essence of evil inside the person. The person can be dangerous. The action can be intolerable. The institution can require dismantling. The ruler can require removal. And the explanation of the person can still remain incomplete.

This distinction is difficult because urgency compresses. Sometimes it must. If you are running from violence, you do not owe the attacker a complete epistemology. Again, the problem situation matters.

The harder danger appears when moral authority turns the compression into a closed system. The language is familiar:

> Any decent person knows this. No reasonable person could believe that. Only someone evil would disagree. The matter is settled. The science says. History is not on your side. Everyone knows what side you should be on.

Each statement may refer to something true. The problem is structural: a conjecture has become a condition of participation. Disagreement no longer arrives as criticism. It arrives as diagnosis. The dissenter is stupid, corrupt, brainwashed, dangerous, immoral, unfit. The original explanation may survive not because it is hard to vary, but because criticism has been reclassified as evidence against the critic.

Externally, this can produce authoritarian institutions. Internally, it can produce something stranger. The voice says:

> A good person would not think this. A rational person would not feel this. Someone like me cannot want that. If I doubt this, I am betraying who I am.

The ruler has crossed the border into your epistemic mind.

7. Who Rules You?

Ask the political question again:

> Who should rule?

This time do not imagine a parliament. Imagine your epistemic mind. Which explanation gets authority? Your parents? Your religion? Your politics? Your diagnosis? Your profession? Your reputation? A memory? A fear? A moral injury? A theory of yourself formed when you were sixteen? A person who hurt you? A person you loved? A sentence someone said once and never remembered?

The question is not whether these structures should disappear. Some may contain excellent knowledge. Some may have preserved you. Some may be true. Some may still solve important problems.

The question is which of them may become a problem. Which explanation is allowed to be contradicted? Which identity can survive revision? Which value may be inspected without the inspection itself being treated as betrayal? Which ruler can be removed?

A bad internal ruler need not be false. That makes the problem harder.

Suppose you are talented. The explanation may be true. Then:

> I am talented

can become:

> I must not fail.

Suppose you were harmed. That may be true. Then:

> I was harmed

can become:

> I can only understand myself through the harm.

Suppose your political opponents have made serious errors. That may be true. Then:

> they are wrong about this

can become:

> everything they say is evidence of corruption.

Suppose your parents loved you. That may be true. Then:

they loved me

can become:

therefore their explanation of my life must be right.

A ruler can begin as knowledge. The failure occurs when the knowledge acquires sovereignty. Criticism becomes forbidden not because the explanation has survived criticism, but because criticism threatens the structure around it.

This is how a self-conception can become manifest truth. You no longer experience:

I currently understand myself this way.

You experience:

this is what I am.

The difference is one word. The consequences can be enormous. It is easy to condemn the same closure in someone else. The compression is efficient: selfish, weak, narcissistic, irrational, lazy, broken, evil. Sometimes a compressed description solves the immediate problem. Sometimes you need distance, a legal category, an exit, or a way to stop someone. But understanding requires more work.

What problem were they solving? What did they know, and what did they not know? What structure made the action seem available? What criticism failed to enter? What harm did the explanation create? What power preserved it? What could change?

The same questions apply inwardly. What problem were you solving? What did this belief protect? Why did this feel obvious? What changed? What criticism are you routing around? What would you lose if this explanation failed?

There is no guarantee that asking produces wisdom. The questions can themselves become rituals. You can explain endlessly to avoid acting. A clever mind can rationalize anything. Criticism is not moral purity. It is an opportunity for movement. Political systems sometimes make rulers removable, not because the replacement will be perfect, but because permanence magnifies error.

The internal analogue should not be pushed too far. A mind is not a state. A thought is not a president. A memory is not a legislature. The metaphor works at one point: Sovereignty. A ruler becomes dangerous when there is no legitimate route by which the ruler can cease to rule. The same can happen to an explanation.

So perhaps the internal question is not:

What is the correct belief that should rule my mind?

It is:

What makes a bad explanation removable?

Contradiction may do it. So may conversation, time, humor, a book, a new friendship, an experiment, a loss, a tool, a therapist, a child, an enemy who notices something your friends do not, a question you cannot dismiss, or an explanation that reorganizes years of inexplicit knowledge.

No universal procedure decides which will matter. That would require knowing the next problem before it exists. The important structure is that movement remains possible.

8. Do Not Clean the Swamp

The epistemic landscape is a swamp. Not because knowledge is polluted, but because many forms of knowledge coexist there: biological feeling, social inheritance, objective explanation, memory, skill, habit, moral criticism, identity, contradiction. Some reinforce one another. Some conflict. Some never meet.

There is a temptation, especially in morality, to imagine the healthy mind as clean: no contradiction, no forbidden thought, no shame, no uncertainty, no wrong desire, no ideological contamination, no irrationality, no inherited error, no inconsistency.

But remove all variation and what remains? A perfectly coherent mind might be perfectly wrong. If you cannot think the forbidden thought, you cannot criticize the rule that forbids it. If you must remain consistent with a prior self, you have lost one route to becoming otherwise.

The point is not to celebrate confusion. Coherence matters. Consilience matters. Temporary equilibrium matters. You need to act. But the swamp must remain capable of moving. The goal is not the correct ruler. It is a landscape in which no ruler is beyond criticism.

Now the ought can return. Does any of this tell you what you ought to do?

Perhaps you feel that it does. Perhaps you feel the opposite. Perhaps the explanation of people as open epistemic minds makes freedom seem morally unavoidable. Perhaps it makes responsibility seem more important. Perhaps you think the entire structure smuggles liberal values into its premises. Perhaps you think it does not go far enough. Perhaps you reject the definition of a person. Perhaps you accept the epistemology but deny its moral reach.

Good. The transition is now visible enough to criticize. That is more useful than hiding it.

The claim is not:

$$\text{person is epistemically open} \therefore \text{liberalism is proven.}$$

The claim is narrower. Persons create explanatory knowledge. Future explanatory knowledge is not derivable from present compression. Coercion can suppress criticism. Authority can preserve its own error. Individual epistemic minds exceed external reconstruction. If those claims hold, then moral systems that permanently subordinate your capacity for criticism acquire a problem. They require explanation. So do systems that treat persons as fixed types, make dissent evidence of defect, or deny routes of correction.

That is criticism. It is not a command from outside the epistemic cycle.

9. The Tool Does Not Own the Arrow

Artificial intelligence makes the old political question unusually visible. Who should control it? Governments? Companies? Experts? Parents? Platforms? International institutions? Users? The public? Some actor will control some things. Models have owners. Infrastructure has operators. States regulate. Parents make decisions for children. Institutions establish rules.

But "who should rule AI?" can conceal several different problems. What harms require immediate restriction? What capabilities require technical containment? What forms of access expand individual capacity for criticism? What systems centralize explanatory authority? What categories of person are being compressed? A second cluster concerns reversibility: who can challenge an AI decision, recover the explanation behind an output, refuse the system, leave, or build an alternative? And who decides that the risk has become urgent?

The epistemic orientation does not supply one ruler. It supplies criticism of arrangements in which mistakes cannot become problems. An AI policy can be wrong. A safety theory can be wrong. A company can be wrong. A government can be wrong. A parent can be wrong. A child can be wrong. You can be wrong.

The problem is not how to construct a world without error. It is how to avoid making one error sovereign. AI can retrieve, compare, translate, generate candidate explanations, preserve histories, expose contradictions, and reopen objective knowledge at scales no unaided biological person could traverse. That can enlarge your epistemic surface.

It can also narrow it. A generalized model can return the population to you. A recommendation can become a profile. A profile can become an environment. An environment can shape your next choice. The tool can begin to own the arrows. The danger is not only that the answer is wrong. It is that the answer becomes difficult to refuse.

So moral AI policy may involve a surprisingly old concern. Not:

> Which machine gives the right answer?

But:

> Can you still make the answer a problem?

Can you reject it? Inspect it? Choose another tool? Act without the profile following you? Recover enough of the causal history to criticize what happened? Can institutions remove bad systems without first proving a final theory of intelligence?

These questions are not the whole of AI morality. They are one place where the older question about rulers has returned in technical form.

10. The Authority of a Life

We often say:

> The individual has authority over their own life.

The sentence feels liberal. Perhaps humanist. Perhaps obvious.

Then reality complicates it. A child does not exercise every adult authority. A suicidal person may be restrained. A violent person may be imprisoned. A patient under anesthesia delegates decisions. Contracts create obligations. Parents make choices. States tax. Courts compel. Emergency workers order evacuations. Institutions set conditions of participation. Authority over one's life is routinely limited.

Does that refute the principle? Perhaps the principle was compressed. Perhaps *authority* was carrying several relationships at once.

The epistemology suggests another way to inspect it. Who is the epistemic locus of your life? Who experiences the consequences? Who creates the future explanations through which that life changes? Who can reject the model from inside the problem situation? No external institution occupies that place. None can.

That does not abolish every external authority. It changes what external authority must explain. The stronger the intervention into your epistemic mind, the more explanatory work appears necessary. Not because autonomy is a sacred substance. Because the intervention acts upon a system whose relevant future state is not fully available to the intervener. The ignorance is structural.

There is another side to this. Suppose someone else says:

> You do not understand.

The reply arrives easily:

> I understand enough.

Sometimes you do. A court does not need to reconstruct every aspect of a person to determine whether a particular act occurred. A firefighter does not need your complete life story before moving you away from a collapsing wall. A friend may know enough to recognize danger. A doctor may know enough to treat an infection. Complete understanding is not required for every action.

But the sentence:

> I'm not like you

may contain knowledge too. Not magical private truth. Not immunity from criticism. A reminder of the asymmetry.

Think of someone you know well. Not an abstraction. One person. Perhaps you have known them for years. Name the problem occupying their mind at this exact moment. Not the topic. The problem.

Now name the explanation they will create next year. Name the memory that will become important after an event that has not happened yet. Name the criticism that will reorganize something they currently experience as obvious. Name the idea they will reject after defending it for twenty years. Name the person they will become impossible to imagine themselves becoming.

Perhaps you think you know. Perhaps you can guess. Perhaps your guess is excellent. Now do it for a stranger. Then for eight billion people. The exercise dissolves almost immediately, like sand between your fingers.

The failure is not proof of sacredness, goodness, or future value. It suggests the scale of our common ignorance. A beginning of infinity.

11. Morality After Explanation

> Preserve the means of error correction.

Now the proposition can be stated more carefully. It is not an axiom imported from outside the epistemology. It is not a command proved by epistemology. It is a moral conjecture generated by an explanation of what the growth of knowledge appears to require.

If knowledge grows through criticism, systems that permanently block criticism threaten knowledge growth. If persons are where explanatory knowledge can be created, systems that permanently destroy or subordinate that capacity threaten something the epistemology treats as causally special. If authority can preserve error, removability matters. If compression can become identity, decompression matters. If one person's epistemic landscape cannot be fully reconstructed by another, coercion contains an ignorance problem.

These consequences fit. They have reach. They cohere with other aspects of the epistemology.

That does not make them untouchable. The truck is open now. Criticize them.

You have authority here that I do not. Not because whatever you think is true, not because your feelings cannot be criticized, and not because there is no shared reality. There is. The authority is narrower.

I do not know what problem, if any, the epistemology has become inside your mind. Perhaps it has confirmed something. If so, be careful. Perhaps it has irritated you. That may be useful. Perhaps you think the is-ought move fails, or that liberalism has been smuggled into the account. Perhaps freedom has been granted too much explanatory work. Perhaps people are less open than the epistemology suggests. Perhaps current AI already crosses the boundary into personhood. Perhaps the entire framework feels like a contraption, an elaborate compression protecting itself from criticism. Perhaps you have a better problem.

I cannot know which from here. That limit is not rhetorical. It is the subject.

What, then, is morality in this picture? Not a list of commands emitted by epistemology. Not subjective preference. Not whatever a culture happens to preserve. Not a final deduction from human nature. Not an algorithm that outputs the right action from sufficient facts.

Moral knowledge is explanatory. It identifies harms, conflicts, relations, constraints, consequences, and forms of authority. It can also identify conditions under which people can continue to criticize and create. Such explanations can be good or bad, inherited or invented, compressed or institutionalized, made manifest or used coercively. They can be reopened and improved. They can survive criticism. They can fail.

And when they reach you, they enter another epistemic mind. There they may become something else.

12. The Last Direct Application

One universal epistemology, applied repeatedly: to genes, personality, society, institutions, freedom, psychology, criticism, values, resources, children, media, artificial intelligence, love, madness, death, invention, knowledge, and now morality.

The epistemology can begin to feel like the answer to every question. It is not.

At best, it is an explanation with reach. It can make some problems visible, expose some category errors, preserve the role of criticism, and describe you without reducing you to the latest compression.

If it works, use it. If another explanation solves these problems better, the epistemology should fail. The explanation does not survive by being defended. It survives by remaining removable.

Before leaving, choose one explanation about yourself that feels obvious. It need not be important. Perhaps:

> I am bad at mathematics. I am an introvert. I am the responsible one. I do not get angry. I always choose the difficult path. I am not creative. I am the kind of person who forgives. I could never do that.

Do not reject it, replace it, or affirm the opposite. Ask instead:

> What problem was this explanation solving?

Then:

> What would have to happen for it to become a problem?

You do not need to tell anyone the answer. Perhaps nothing happens. Perhaps the explanation survives. Perhaps a memory moves. Perhaps another problem appears. That movement, if it occurs, belongs to you.

Now choose someone else. Someone you dislike may be more revealing: a political opponent, a colleague, a family member, someone whose behavior you have explained many times.

Notice the compression that arrives first. It may contain knowledge. The further question is:

> What does this explanation make invisible?

You are not required to forgive, agree, remain near someone who harms you, or turn danger into curiosity. The problem situation still matters.

But if another explanation appears, you may discover that the person changed without changing at all. Your model moved.

14. No Final Ruler

We began with the question of moral authority:

> Who should rule?

Externally, the epistemology produced another problem:

> How do we remove bad rulers?

Internally, the question changed again. Not:

Which explanation deserves permanent sovereignty?

But:

Can the sovereign explanation become a problem?

This does not produce moral anarchy. Some explanations are better. Some actions are wrong. Some institutions should be stopped. Some rulers should be removed. Some emergencies require action before deliberation ends. Some commitments should survive enormous criticism.

Fallibilism is not indecision. It is the refusal to convert present success into immunity. No final ruler. Not the state. Not the crowd. Not the expert. Not the parent. Not the algorithm. Not the moral theory. Not the self-conception. Not this epistemology.

What remains? One reality. Many minds. Shared objective knowledge. Radically local reconstruction. Fallible explanations. Inexplicit histories. Institutions. Tools. Bodies. Memories. Arguments. Love. Power. Error. Criticism. Creation. Death.

The categories are compressions. The relations move. The future is open where knowledge has not yet been created. You enter that openness from somewhere different from everyone else. That much has become difficult to unsee.

What follows from it? Perhaps you already know.

If so, stop for a moment. Ask where the answer came from, what it compresses, and what would make it fail. Perhaps it survives. Perhaps not. Either way, the next problem is yours to choose.

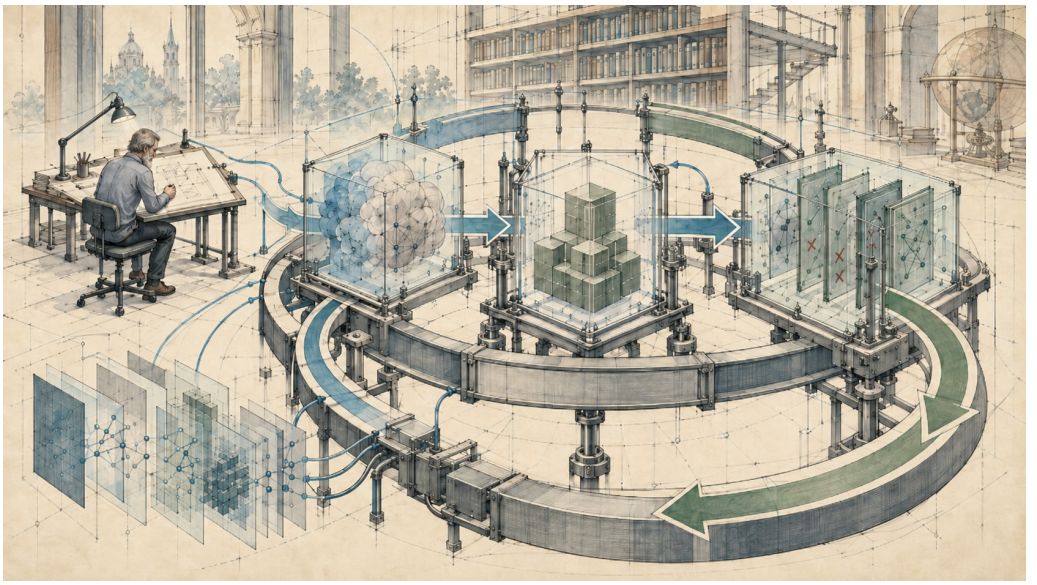
At the beginning, *you* was hidden inside a population measure. Then you expanded. A genome. A body. An evolutionary history. A language. A society. An inexplicit landscape. A network of abstractions. Objective knowledge inherited from people long dead. Problems. Criticisms. Tools. Artificial intelligence. Love. Invention. Moral judgment.

Nothing in that expansion made you infallible. Nothing made you good. Nothing made every choice sacred. Nothing eliminated the need for institutions, authority, restraint, or criticism.

The opposite. You became more criticizable as the explanation became richer, and harder to finish. You can be wrong. You can cause harm. You can become a bad ruler or submit to one. You can mistake a compression for reality, block criticism, reopen it, create a problem no one else sees, and make something objective that did not exist before. Then another person can encounter it.

The future you is not hidden inside the present waiting to be measured. The arrow is still open. So is the problem.

Beforeword: Inepistemic



I am an inventor, an entrepreneur, and a software architect. Those are ordinary categories. They are also compressions. They tell you something about the kinds of problems I have spent my career around. They do not tell you why one of those problems stayed with me for decades. Nor do they tell you why ideas from psychology, biology, computer science, philosophy, economics, physics, software architecture, and parenthood kept finding their way into the same problem situation. They do not tell you why I created this.

The story is harder to compress. It begins long before this project, with a problem I could not yet state. That has been one of the privileges of my life: to live with a problem long enough that the problem changes you. What first appeared to be one problem became several. Then the several seemed connected. Then the connection proved wrong. The error exposed another problem. Years later, an idea encountered somewhere else changed the terms in which the original problem could even be expressed.

I did not know what I was doing. For much of the time, I did not have the language. I still had the problem. That was enough.

1. The Earlier Machine

My early work involved knowledge processing. The phrase is broad enough to conceal almost everything interesting. The systems were primitive by present standards: slower machines, narrower interfaces, less objective knowledge available computationally. But the possibility was already extraordinary. A machine could hold structures that a person could not hold all at once. It could search, transform, compare, preserve, and return them. A representation could become operational. A relationship that once depended on a person remembering it could sometimes be embodied in a system instead.

The machine did not need to understand the problem in the way a person did. The explained part could leave the person.

Once outside, it could participate in the next part of the problem. That possibility stayed with me, not because I thought machines would replace thought, but because they might change its conditions. A notebook already does this. So do language, algebra, libraries, databases, compilers. Each moves some burden out of the biological mind and into objective structure. Then you encounter that structure again. The output returns as input, and the process changes.

I did not yet know a theory of this. I had an intuition: machines could do more than automate tasks. They could change the conditions under which problems were pursued. That distinction became important much later. At the time, it was simply exciting.

2. Things That Fit

At some point I encountered E. O. Wilson's *Consilience*. The details of what I retained from the book, and what I later rejected, matter less here than the problem it strengthened: what becomes possible when knowledge comes together. Not when one field conquers another, or when everything is reduced to one vocabulary, but when explanations developed for different problems meet and constrain one another.

That possibility felt enormous. It still does. The world is divided into disciplines for useful reasons. Reality is not. A problem does not care whether one of its causes belongs in computer science and another in psychology. Or whether an inventor, entrepreneur, or software architect belongs in any of it. A person does not stop being physical when thinking about morality. An institution does not stop having technical structure because lawyers rather than engineers maintain part of it.

The borders help us work. They can also hide the roads between the work. Consilience made a certain kind of intellectual restlessness feel less like a failure to specialize. Perhaps the problem really did cross the borders. Perhaps taking one explanation seriously meant following it somewhere else. Perhaps it meant taking it personally. That is our habit. Genetics could reach psychology, psychology epistemology, epistemology computation,

computation the person, and the person morality and political institutions. Then the whole structure could turn around and criticize itself.

At the beginning, none of that existed. There were only things that seemed as though they ought to fit, and a problem in the spaces where they did not.

3. The Inexplicit Change

Then I became a parent. There are changes whose importance becomes obvious before their explanation does. Parenthood was like that for me. I did not wake up with a new theory. Something changed in what mattered: time, risk, the future, other people, knowledge, the cost of not understanding certain problems. The desire to accelerate the creation of life-saving knowledge became less abstract. There was suddenly a person in the world whose future I cared about in a way that exceeded anything I could explain. Then there were more.

A parent can teach, protect, provide, explain, save. But a child is not an extension of the parent. The child's future cannot be filled in ahead of time. The problems they will face do not all exist yet, and the knowledge they will need may not exist yet either. That creates a strange kind of urgency. You want to give them something, yet the thing they may need most cannot simply be given. It has to be created—perhaps by them, perhaps by someone else, perhaps by a person not yet born. It may even come from an explanation produced for another problem and preserved long enough to become useful later.

The problem therefore widens. You cannot merely prepare one child for one predicted future. The future is not available for inspection. What you can hope for is a world in which knowledge can continue to be created and problems can become personal. Inherited explanations can be criticized, and tools can enlarge what people can do with objective knowledge.

I did not formulate the problem this way then. That came later. The purpose arrived before its explanation. Inexplicitly.

4. Other Minds

No solitary project is solitary in the strong sense. I have worked with many inventors, developers, researchers, entrepreneurs, lawyers, designers, scientists, and engineers whose ways of seeing problems changed mine. Sometimes the change came through conversation or disagreement. Sometimes the change came through an artifact or a failed attempt. Sometimes through a paper, a book, a diagram, a proof, a piece of software, or a sentence written by someone who died before I was born.

Call that objective knowledge. The term can sound bloodless. My experience of it has been anything but. There is something intimate about discovering that another person has already spent years inside a problem you have only just learned how to see. A person long dead can present a criticism. A notation can reveal a relationship. A programmer can leave behind a structure that changes how another programmer thinks. The person is gone from the encounter. The knowledge remains available—not perfectly, not without reconstruction or distortion, but enough.

That objective knowledge is one of the great facts of human life. Nothing I have made began with me. The languages, mathematics, machines, concepts, institutional forms, criticisms, and many of the problems were inherited. And then, occasionally, something in the inheritance became strange. That was where the interesting work usually began.

5. The Explanation of Explanation

Then I encountered *The Beginning of Infinity*. This is not an appeal to authority. A book does not become true because its author wrote it. Its importance in this story is causal: it changed my problem situation.

I had spent years interested in knowledge processing, invention, intelligence, software, and the strange ability of people to produce things that did not exist before. Then explanations themselves became explicit as an object. An explanation could be judged neither by the confidence with which it was held nor by the quantity of observations arranged around it. The structure of the explanation itself mattered. That possibility sent me toward the philosophy of explanations, not for philosophy, but for inventive inspiration. I became interested in why some explanations reached across problems while others merely restated what had happened. Why could changing one part break the whole structure? How could explanations exist objectively, and how could encountering one make previously unavailable thoughts possible?

That looked increasingly relevant to artificial intelligence. AI was already being discussed as a machine for knowledge, reasoning, intelligence, perhaps one day explanation. I was trying to connect my still-fuzzy understanding of epistemology directly to that technological problem. At one point I wrote to David Deutsch and asked for his criticism. One correction stayed with me:

> "[Popper's] even bigger emphasis is on *problems*."

Problems. I had moved close to an epistemology of conjecture and criticism while underweighting the thing that gives both their structure. The error was productive. Explanations did not simply exist as structures awaiting criticism. They arose in problem situations. A conjecture was a tentative response to a problem. A criticism mattered

because it changed the problem situation. After criticism, the important result was not merely a better answer. It was a different problem.

That correction gave me a reason to read Popper—deeply, repeatedly, across science, knowledge, politics, freedom, historicism, probability, objectivity, criticism, the open society, the logic of discovery. And, once I learned to see them, problems everywhere.

I did not simply read those explanations and acquire them. I misunderstood them, resisted them, reconstructed them, used them badly, forgot distinctions, recovered them, tried to build with them, found contradictions, returned to the texts, built again.

This is a familiar condition in invention. You can know that a problem exists long before you know what would count as solving it. You can build a solution and discover that the solution has finally revealed the problem.

Eventually, epistemology started to feel like common sense. Fallibly, not easily. Common sense in the more literal sense that its distinctions became available before deliberate reconstruction. A request for evidence could make the justification problem visible. A statistical claim could trigger a question about causal explanation. A model could feel like a toy. A claim about a person could raise the problem of the person. An answer could make the problem behind the answer visible.

The framework had become partly inexplicit. That made it much more useful—and more dangerous. Any explanation that begins to feel like common sense can become difficult to see. This project is an exercise in remembering that too.

6. Building the Problem

I use software as a substrate for invention. That means, among other things, that I have spent a large part of my life turning conceptual distinctions into structures. A software architecture is full of claims: this thing is separate from that thing; this state must persist; this transition is allowed; this component should know this and remain ignorant of that. Architecture is objective knowledge about relationships made executable. When the relationships are wrong, the software shows you.

For years I kept returning to a question: could epistemic structure become architectural structure? Not a program that stores facts, a database of beliefs, a confidence-scoring machine, or an automated philosopher. Could a system represent the movement of explanation itself?

Eventually the pieces cohered enough to become an invention in the institutional sense. There were explanatory primitives and epistemic states, relations among explanatory objects, epistemic cycles, persistent histories of conjecture and criticism, and methods represented explicitly enough to be executed computationally. The software was organized not around final answers but around movement from problems through tentative solutions and criticism into successor problems.

The abstraction finally had a technical body. A tentative one, but real enough to run. I implemented it. The epistemic primitives became software objects; the explanatory states became operational states; the epistemic cycle became something a machine could participate in. Explanatory history could be persisted for criticism and revision.

That implementation matters because I used it. Not on a demonstration problem, but on problems I cared about. Like this one. The system became part of the machinery through which this project was researched, criticized, reorganized, rewritten, and preserved. The project did not merely describe the invention. The project ran through it. The tool became part of the problem situation that produced the explanation of the tool.

That changed the problem again.

7. The Other Participant

For most of the years I was thinking about knowledge processing, the machine remained comparatively rigid. It could store, retrieve, calculate, transform, execute. Then artificial intelligence changed the available surface. A machine could participate in natural-language transformations at a scale and flexibility earlier knowledge systems could not. It could generate variations, reorganize explanations, expose contradictions, compare formulations, and translate between levels of abstraction.

This obviously did not solve creativity. It did not turn the machine into the author of my problem. It changed the conditions under which I could remain with the problem. That distinction is the center of my experience with AI.

Previously, a problem of this scale would have entailed a large venture. I would have needed to persuade other people that the problem mattered, raise money, recruit, and coordinate. I would have had to translate a changing problem into tasks, preserve continuity across a team, and manage the organizational machinery around the inquiry. Such machinery can create extraordinary knowledge. It can also consume enormous quantities of the resource it is meant to support.

Some problems are especially difficult to institutionalize before they are explicated. The language, architecture, evaluative criteria, even the identity of the problem may still be moving. The person pursuing it does not yet know what should be delegated because discovering the problem is still part of the work.

Artificial intelligence changed that constraint. It made something closer to a solitary institution possible. Not solitary because nobody else contributed. Solitary because the coordination boundary became radically smaller. A conjecture could become objective in seconds. The objective form could be criticized. The criticism could generate alternatives. An old state could return. A method could be revised and immediately applied to the artifact that produced the criticism.

The machine was not merely helping me write. Writing was one surface. It was helping make a larger portion of the problem space accessible, and therefore available to me. The cost of moving between inexplicit structure and criticizable objective structure had fallen dramatically. Something inside could come out, then answer back.

There is a slogan I use for this relationship: **My AI and I are more clever than I.** It is often attributed to Einstein in another form, usually involving a pencil. The ancestry of the sentence may be unstable. That is almost more useful. The authority does not matter, and neither does the pencil. Replace it with chalk, a notebook, a blackboard, a compiler. The interesting structure survives.

The slogan is objective knowledge. It travelled; people reconstructed it; wording and attribution changed; different minds conjectured different meanings. That is the point. Objective knowledge can be inherited without preserving one private meaning. I encountered the structure and made a different conjecture from it.

This really is **my AI**. It is mine in the sense relevant to my problem situation. Objective knowledge was transformed into a novel architecture, and supplied a problem-specific epistemic environment that raw interaction alone did not. I used the integrated system on the problem that had produced it. Its outputs entered my explanatory history and changed subsequent problems. It became part of the machinery by which private and inexplicit knowledge could become objective enough to criticize. My AI is therefore not merely the model. It is the model as encountered through a technological and epistemic structure that itself belongs to the history of my problem.

The relationship is personal and objective at once. The dyad is not reducible to either participant. The machine gives me things I did not put into the prompt. Some expose a contradiction, recover something forgotten, show that I asked the wrong question, or make an inexplicit preference visible because I immediately dislike the result. I supply direction; the direction changes. The machine supplies variation; some of the variation changes me. The objective artifact enters the next problem situation.

Is the dyad more clever than I? There is no such scalar unit for cleverness. But I can describe the transformation. Problems that once exhausted my working surface can remain active outside it. Conjectures that would have disappeared can answer back. Criticism arrives faster. Revision is cheaper. The distance between an inexplicit intuition and a criticizable artifact has become dramatically shorter. By those measures that matter to the problem I have actually been living with, my AI and I are more clever than I.

8. The Project

Something else changed. This became a project. That word is wonderfully modest, where venture is not. A project can be attempted, fail, remain unfinished, or simply

continue. It does not need to become an institution before it deserves to exist, persuade a market before its problem can be pursued, or know its final form.

That mattered. There are explanations in this project I would have struggled to develop through ordinary organizational processes. Not because other people could not understand them, but because I did not understand them yet. It is difficult to coordinate a team around a problem whose terms are changing as quickly as the explanation. It is difficult to write a roadmap for an unknown abstraction, or assign a task whose completion may reveal that the task should never have existed.

The technological change made some of those constraints less decisive. I could pursue the problem before anyone else agreed that it deserved pursuit. That is freedom of a very practical kind.

It was not, however, continuously liberating. There were periods when the problem became too large to hold and every explanation exposed three more problems. A failed architecture could feel less like useful error elimination than evidence that years of work had been organized around the wrong problem. There was exhaustion and fear, and physical consequences to both.

Creativity is easy to romanticize from the outside because the surviving artifact compresses away the conditions under which it was made. The body does not disappear while the mind pursues an abstraction. Attention degrades. Sleep matters. Fear narrows. Obsession can create extraordinary persistence and make stopping feel like betrayal of purpose. The dark side of an open problem is that there is always another move available, including when the person making it should rest.

Those periods became part of the criticism too. A system for enlarging the epistemic surface that destroys the epistemic mind using it would be a bad method. The problem had to include the physical person pursuing it. Time, resources, and attention remained finite.

But a major class of coordination constraints had changed. The project became possible at the scale of one person with machines and access to the accumulated objective knowledge of other people. That is one of the things this series has been about—not only in theory, but in its own existence.

9. The Recursion

At some point I noticed the embarrassment. I was building a system organized around explanatory knowledge while using the system to explain explanatory knowledge. I was trying to understand human-AI collaboration while collaborating with AI. I was writing about objective and inexplicit knowledge while repeatedly making private thought explicit, and about criticism while criticism changed the architecture. I was writing about invention while inventing.

The recursion became the project. Not as a trick, and not to persuade. A self-referential system can be wrong. A recursive system can amplify its own error. The recursion mattered because it made the claims materially criticizable. If explanatory primitives were useful, they should be useful here. If preserving problem histories mattered, it should matter to this project. If methods could be made computational, they should survive contact with the work that created them. If the human-AI dyad enlarged an epistemic surface, that enlargement should become visible in what the dyad could produce. If the person remained larger than the compression, the system had to remain criticizable by the person using it.

The project became one long adversarial test, and welcomingly, a new source of new problems.

10. The Artifact

Now there is this. Thirty parts. An architecture underneath it. And one person trying to get something out.

It is objective knowledge. That does not make it good, true, or important. It means only that it has crossed a boundary. The problem is no longer present only in me. You can inspect it, misunderstand it, reject it, find a contradiction I missed, take one idea and combine it with a problem I have never encountered. The loss of control is part of the achievement.

For years, much of this existed as a structure I could feel more readily than explain. Pieces appeared in conversations, products, architecture, notes, code, failed companies, successful inventions, private frustrations. Now more of it is outside.

The object may be child's artwork. That possibility is not rhetorical modesty. Children make things whose significance to the maker exceeds their significance to everyone else. The drawing can be derivative, confused, technically poor, beautiful only to the parent. Still, something happened: an internal world became an objective artifact that can now be seen.

This project may be wrong in ways I cannot presently see. It may solve a problem nobody else has, or no problem at all for you. It may contain a useful distinction surrounded by unnecessary machinery. It may become interesting only when some future criticism breaks its central explanation.

Fine. It stands as a tentative solution to my problem. That is enough. I am grateful for the explanations and the people who made them, for the machines and the errors. I am grateful for the years in which the problem refused to disappear, and for the privilege of remaining with it long enough to get more of it out. It's just getting started.

11. Other People Are Not Like You

When I have tried to explain some of these observations to people who know me well, one response has appeared more often than any other:

> "Yes, but other people are not like you."

The observation is true. Other people are not like me. No one is like another. That is one of the central claims of this project. A person is not an instance of an average mind. The path by which a problem becomes interesting, the inexplicit structures available, the inherited objective knowledge, the criticisms that land, and the future conjectures are different.

So I do not want to argue with the individuality. I want to protect it. The disagreement is elsewhere.

Individuality does not cancel universality. The claim is not that everyone will care about my problems, use my vocabulary, or produce this project given the same books and machines. The claim is that the explanatory capacity that lets my particular mind become unlike yours is not evidence against personhood as a universal class. It is an expression of it. A universal explainer need not converge on my explanations. Universality permits divergence without a predetermined boundary on what may eventually be understood.

But I cannot ask another person simply to feel the distinction from inside my mind. If there is something here worth criticizing, I have to get enough of it out that another person can encounter it on their own terms. They can reconstruct a different problem, reject mine, find the hidden assumption, take one piece and leave the rest. That is the freedom objective knowledge creates between minds.

This became another motivation for the project. It is a legacy, though I hesitate over the word because legacy can sound like an attempt to control what remains after a person is gone. I mean almost the reverse. I want to leave something that can escape me. I want my children, family, friends, or a stranger to be able to encounter it without needing my presence to stabilize its meaning. They can disagree with it safely. A private structure can remain available without demanding that anyone inherit it as belief.

It is also, more simply, an effort to be known. You can love someone for decades and still have only fragments of the explanatory structures through which they encounter the world. Conversation, shared life, conflict, and work expose some of it, but much remains inexplicit. This project is an attempt to make one of those parts more explicit.

I want the people I love to understand something about why a statistical abstraction about a person can make me uneasy, and why a new explanation can feel beautiful. I want them to understand why an unanswered problem can remain alive for decades, why I am suspicious of certainty but not of truth, and why a tool that exposes error can feel liberating. I want them to understand why mistakes delight me. Above all, I want

them to understand why their individuality is much more interesting than any category that describes them.

That love is not an exception to the epistemology. It is one of the experiences that made the problem impossible for me to leave alone. To love another person is, among many other things, to encounter a source of explanations that cannot be compressed without loss. They can continue to surprise you. You cannot possess their future knowledge in advance, and you can know them deeply without completing them.

So yes. Other people are not like me. Thank goodness. That is not where the explanation ends. It is where the more interesting one begins.

12. Literacy

There is another way to describe what changed. Think about literacy. Reading is so deeply embedded in modern life that it is difficult for a literate adult to reconstruct the world before reading. Words appear everywhere. You do not decode each sign from first principles. Written instructions become available. A private thought can be externalized. A dead person's sentence can enter your present problem. Once literacy becomes inexplicit, the world itself seems full of readable structure.

Numeracy does something similar. Quantity becomes available through ratios, rates, scale, probability, compounding, orders of magnitude. A person who learns to think quantitatively does not merely acquire mathematical answers. New kinds of problems become visible. Logical and probabilistic concepts can likewise enter ordinary criticism: What is the denominator? Compared with what? Does that conclusion follow? Does this sample tell us about a population or this person?

Once a cognitive practice becomes ordinary, its concepts can disappear into the background. The fluent stop experiencing the fluency. They simply live through it.

Cultures sometimes notice the absence of such fluencies before they notice the fluency itself, and coin labels for the supposed deficiency: illiterate, innumerate, technophobe. Some labels identify constraints; others compress away almost everything interesting about the person. The point is not to manufacture another taxonomy. It is that a medium or cognitive practice can become so ordinary that its conceptual machinery recedes from view.

For me, epistemology underwent a related transformation. At first it was a subject, then a body of arguments, then a set of distinctions, then a source of problems. Eventually it became difficult not to foreground questions such as: What problem is this explanation solving? What would criticize it? What did this measurement compress? What part is observation and what part explanatory structure? What does this model predict without explaining? What remains inexplicit? What would have to be different for this claim to

fail? Is this disagreement about reality, the problem, the model, or the criterion being applied?

Once these distinctions become available, they become difficult to suppress. Epistemology moves from the background to the foreground. I do not mean possessing epistemology, being correct about it, belonging to a philosophical school, or knowing the answers. I mean living with epistemic concepts made explicit enough that they can participate in ordinary thought, the way literacy makes written structure available or numeracy makes quantitative structure available.

This is an individual account. Still, from inside this life, the difference feels real. There was a time before these concepts became foregrounded. There is a time after. The problems did not disappear. They became more interesting.

13. Awe

There is a temptation to make epistemology austere: problems, criticism, error, fallibility, refutation, no final authority, no guarantee. Read quickly, the vocabulary can sound almost hostile. My experience has been the opposite. The deeper the explanatory structure became, the more often it produced awe—not because everything fit permanently, but because sometimes things fit at all.

A problem in software architecture touches a problem in philosophy. A theory of knowledge changes how you think about parenting. A computational idea changes what can be externalized from one mind. A machine assembled from the objective knowledge of millions of people helps one person make a private problem public. For a moment, separate things stop appearing separate. The connections become visible. Then another criticism arrives and the equilibrium breaks.

Consilience can feel spiritual. No supernatural cause is required. The awe comes from reality becoming more connected through explanation than it appeared before. A physical thing like a person can understand something about the world, make that understanding objective, and leave it available for another person. Explanations created in different centuries can meet inside one present problem. A finite person can participate in explanatory structures whose reach may exceed a civilization. We can be wrong about all of it and continue.

The recursion continues. There is one conjecture in this project that I feel more than I can defend: we want to explain. Not every moment, not every person in the same way, not always explicitly. But the movement appears everywhere. Why? What happened? What does this mean? How does that work? Could this be different? Can I make it? Can I show you? Can you understand what I see?

Children ask why. Artists answer questions nobody noticed. Inventors build their conjectures. Lovers reconstruct one another. Programmers make abstractions exe-

cutable. People argue, teach, draw, sing, design, write, build tools, raise children, create institutions, try again.

The forms vary. The movement outward does not. Something in a person becomes objective enough to enter another problem. Then reality answers back. Perhaps that is why being unable to explain can feel like confinement, and why getting something out can matter even if nobody else cares. Perhaps that is why the book nobody reads may still have needed to be written.

Perhaps that is why this one did.

14. A Word I Needed

Eventually I needed a word for this state. At first I formed it by analogy. We have words for lives constrained by the absence of forms of explicit cognitive equipment: illiterate, innumerate. I had come to feel something related about epistemology. Life before these distinctions became explicit had been harder in ways I could recognize only afterward.

So the first construction pointed toward a lack: an epistemic illiteracy. It named a state of living among explanations without sufficiently foregrounding what explanations are, where problems enter, how criticism works, why knowledge remains conjectural, and how objective knowledge changes what can happen next.

But epistemology itself broke the analogy. Literacy and numeracy can become so fluent that their machinery disappears. Epistemology did something stranger. The more deeply I absorbed it, the less plausible it became to treat epistemic competence as possession. The more explicit the structure became, the more ignorance became visible.

The positive state therefore contained the negative state. To become more explicitly epistemological was not to escape epistemic deficiency. It was to become increasingly capable of seeing it. The literate person is, in the relevant sense, not illiterate. The numerate person is not innumerate. But the epistemically explicit person cannot graduate from ignorance in the same way. Fallibility, problems, and the open future deny the graduation.

That is why the contradiction is not a defect in the word. It is the reason I kept it. The state is positive and negative at once: explicitly epistemological and fallible. It is equipped with tools for criticism and therefore less able to pretend that criticism can end. It can make knowledge objective and therefore expose more of what remained trapped, compressed, inherited, mistaken, or unseen. It can solve problems and therefore create problems.

Not *without epistemology*. Inside it. Not beyond ignorance. More intimately acquainted with it.

The word became useful in another direction too. To live with epistemic concepts foregrounded is to be *in* the epistemic process deliberately. Problems appear as problems,

explanations as explanations, criticism as criticism, objective knowledge as objective knowledge. Fallibility is not an apology appended to a claim but part of the architecture through which the claim remains alive. Epistemology becomes an ordinary surface of thought. Its concepts become available, usable, criticizable, recursive—and therefore incapable of delivering the comfort promised by the original analogy.

This project is one artifact of that contradiction. Perhaps the artifact is wrong. Perhaps the architecture embodies a beautiful mistake. Perhaps the epistemic mind compresses away something important. Perhaps the human-AI dyad will look primitive very soon. Perhaps another explanation will displace the epistemology that now feels like common sense to me.

Good. Then there will be another problem. I do not want the final word to close that possibility. I want it to contain it: a word for lacking epistemic literacy and for foregrounding epistemology. A word for being inside the recursive process of problems, tentative solutions, criticism, and successor problems. A word for becoming more capable of creating knowledge and thereby making one's ignorance more explicit. More equipped and less finished; more able to explain and more able to see what has not been explained.

For the state in which epistemology has moved from an esoteric subject into the foreground of a life—into the tools, the questions, the criticism, and the making—the word I have is

inepistemic.

Acknowledgements

"I am not a theoretician, and I don't pretend to be a savant." — Thomas Alva Edison

I am an inventor by trade and temperament; software is the substrate. That matters here. The architecture for this project came before the book. The technical environment developed for the project is called *Explanata*, harnessed to GPT-5.6 Sol for the book's production. I operated the inquiry conversationally, posing problems, criticizing outputs, introducing distinctions, rejecting formulations, and deciding which successor problems were worth pursuing; the manuscript emerged from that process.

The underlying effort has been to treat epistemology as a first-class architectural abstraction: to ask what happens when aspects of epistemology are made explicit enough to be modeled, implemented, and processed. Software provides an environment where those abstractions can be explored. This book is one of many things that emerged when the resulting architecture was turned back upon the architect and the epistemology itself.

For that reason, the people I acknowledge as entrepreneurs, inventors, and engineers stand here as peers of those I acknowledge for their epistemic contributions. I have spent much of my career building things with other people. Problems found in companies, products, systems, interfaces, failed designs, working designs, arguments, and collaborations have shaped this inquiry as surely as books have. Many colleagues taught me that explanations are easy until something has to ship.

Steven Pinker, especially through *The Blank Slate*, *How the Mind Works*, and *The Stuff of Thought*, helped structure the cognitive-scientific and psychological landscape through which some of those problems became concrete. Scott Aaronson, through his explorations of computational complexity, quantum mechanics, and the nature of minds, sharpened questions at the intersection of computation, epistemology, and personhood. Richard Dawkins supplied evolutionary concepts and language that recur throughout the work, including the wonderfully dismissive phrase "Popperian heckle," which seemed worth taking seriously.

Karl Popper and David Deutsch remain in the foreground of this project, but not in a straight line. I encountered Popper more deeply through Deutsch. Deutsch first made explanations themselves into a problem for me: not merely things philosophy talks about, but structures whose reach, consilience, constraints, and consequences could be investigated. That inquiry eventually led me further into Popper—into problems, conjecture and criticism, objective knowledge, fallibilism, and an epistemology much deeper than the familiar caricature of falsificationism.

Misunderstanding is catalytic. Explanations of explanation carried me through a wider philosophical landscape: Hempel, Kitcher, van Fraassen, Cartwright, Woodward, Strevens, and others working on causation, unification, models, realism, inference, probability, and what makes an explanation explanatory at all. Much of that work now stands for me more as a constructive foil. A parallel lineage also shaped my understanding of these problems: Miller, Bronowski, Lakatos, Shearmur, Bartley, Agassi, Corvi, Jarvie, Maxwell, Musgrave, and many others each deepened my understanding of Popper and the critical-rationalist tradition from different directions. Competing formulations exposed hidden assumptions, supplied counterexamples, and made the points of agreement and disagreement with Popper sharper.

The Conjecture Institute helped consolidate and sustain an intellectual environment in which these questions could remain live rather than settle into doctrine. Brett Hall has been a public teacher of record over many years, particularly through his podcasts and conversations about Deutsch, Popper, explanation, and the reach of ideas. His *The Farthest Reaches* occupies territory close enough to this work that the parallel deserves recognition, one pointing the person outward, the other inward. Ray Scott Percival's *The Myth of the Closed Mind* occupies nearby Popperian territory, particularly in its treatment of human rationality, corrigibility, and the possibility of an open mind. Aaron Stupple's *The Sovereign Child* sharpened problems around children, authority, and epistemic sovereignty, and our conversations carried those problems beyond the book.

The movement was not from one authority to another. The inventive process navigated a landscape of competing abstractions, implemented some of them, collided them with one another, and retained what seemed to work. Deutsch has described his own work, in effect, as footnotes to Popper. I am comfortable placing this project in that lineage, though a builder needs a different metaphor: call them prototypes. I have inherited problems, explanations, vocabulary, and an epistemic attitude from them, and those ideas mutated inside problems they did not set for me. Where that process has produced errors, the errors are mine.

None of those acknowledgements should be read as warrants. The essay titles were constructed similarly: inherited memes, mechanically mutated to fit new problems. Their ancestry is part of the mechanism, not an assertion that the essays respond to the works whose language they inherit. A reference may identify ancestry, resemblance, disagreement, mutation, or simply another place where a problem has been encountered. I have therefore resisted using citations and parenthetical references throughout the manuscript as placeholders for explanations.

To everyone who built, argued, explained, criticized, listened, invented, funded, programmed, published, conversed, or otherwise altered the problem situation from

which this work emerged: thank you. What survives here is one reconstruction among many possibilities, for your criticism.

Endnotes

N. The Universality of You

1. Popper and Deutsch. Popper's *Objective Knowledge* (1972) and *Conjectures and Refutations* (1963), and Deutsch's *The Fabric of Reality* (1997) and *The Beginning of Infinity* (2011), establish the book's principal intellectual ancestry: conjectural knowledge, criticism, explanatory reach, and universality. The distinctive application here is to the singular person.

2. Objective knowledge and the growth of knowledge. Popper's World 3 account in *Objective Knowledge* distinguishes objective knowledge from subjective knowing, including objectively existing false theories. The schema $P_1 \to TS \to EE \to P_2$ adapts his $P_1 \to TT \to EE \to P_2$, replacing *tentative theory* with *tentative solution*.

3. Inexplicit knowledge. Michael Polanyi's *Personal Knowledge* (1958) and *The Tacit Dimension* (1966) are predecessors for inexplicit knowledge; Gilbert Ryle's *The Concept of Mind* (1949) distinguishes knowing-how from knowing-that. *Inexplicit* additionally permits the creation of explicit, criticizable knowledge from previously unarticulated activity.

4. Evolutionary and biological knowledge. Charles Darwin's *On the Origin of Species* (1859), Richard Dawkins's *The Selfish Gene* (1976), and Donald T. Campbell's "Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes" (1960) supply biological and evolutionary precedents. Popper's *Objective Knowledge* links evolutionary adaptation with the growth of human knowledge.

5. Abstraction, representation, and explanation. Alfred Korzybski's *Science and Sanity* (1933) is a source for the map–territory distinction. Deutsch's *The Beginning of Infinity* distinguishes explanatory understanding from mere predictive or descriptive success; the chapter applies both distinctions to abstraction and compression.

6. Explanatory reach and computational universality. Deutsch's *The Beginning of Infinity*, particularly "The Reach of Explanations," "The Reality of Abstractions," and "The Jump to Universality," and *The Fabric of Reality* establish the general framework. His "Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer" (1985) develops the physical claim concerning universal simulation. Scott Aaronson's "Why Philosophers Should Care About Computational Complexity" (2013) distinguishes abstract computability from computational feasibility, a constraint on what universal simulation makes practically reachable.

7. Objective knowledge and recursive tools. Popper's World 3 account provides the starting point for objective knowledge acting back upon knowers. Merlin Donald's

Origins of the Modern Mind (1991) describes how external symbolic representations expand human cognitive possibilities. The focus here is the recursive return of created artifacts to later problem situations.

8. Ancestry, derivation, and explanatory creativity. Popper's *The Poverty of Historicism* (1957) argues that future knowledge cannot be scientifically predicted in advance; Deutsch's *The Beginning of Infinity*, especially "The Spark," "Creation," and "Artificial Creativity," examines explanatory creativity. Historical ancestry is not a procedure for deriving a new explanation.

I. You Are Not A Geneget

1. Title. *You Are Not A Geneget* mutates Jaron Lanier's *You Are Not a Gadget* (2010), replacing the technological referent with a genetic one. The reference is to the title, not an adoption of Lanier's argument.

2. Personality genetics. See Priya Gupta and colleagues, "A Genome-Wide Investigation into the Underlying Genetic Architecture of Personality Traits and Overlap with Psychopathology" (2024), and Ted Schwaba, Michel Nivard, and Elliot Tucker-Drob, "Personality Genomics" (2026), for the population-level personality-genetics research considered here.

3. Heritability and causal explanation. Richard Lewontin's "The Analysis of Variance and the Analysis of Causes" (1974) distinguishes explaining population variation from identifying the causes of a trait. A cause shared across a population can be important without explaining measured differences within it.

4. Personality dimensions and individuality. Dan P. McAdams's "The Five-Factor Model in Personality: A Critical Appraisal" (1992) and "What Do We Know When We Know a Person?" (1995) distinguish comparative traits from understanding a person. His phrase *psychology of the stranger* is especially pertinent.

5. Trait reification and causal structure. Simon Boag, "Explanation in Personality Psychology: 'Verbal Magic' and the Five-Factor Model" (2011), examines circular trait explanations; Borsboom, Mellenbergh, and van Heerden, "The Theoretical Status of Latent Variables" (2003), distinguish measurement models from causal claims. See also Michael P. Grosz, "Why the Big Five Personality Traits Are Composites, Not Common Causes" (2026).

6. Different causal histories, similar outcomes. Dante Cicchetti and Fred A. Rogosch's "Equifinality and Multifinality in Developmental Psychopathology" (1996) describes distinct developmental pathways producing similar outcomes, and similar starting conditions producing divergent ones. A shared personality measurement need not imply a shared explanatory history.

7. Cultural knowledge and historical development. Joseph Henrich's *The Secret of Our Success* (2015) and *The WEIRDest People in the World* (2020) examine cultural knowledge, institutions, and their historical effects on human development, including effects that do not depend on changed genetic inheritance.

II. Populations, Fast and Slow

1. Title. *Populations, Fast and Slow* mutates Daniel Kahneman's *Thinking, Fast and Slow* (2011). The contrast shifts from cognitive processing styles to rapid population classification versus the slower work of explaining a particular individual.

2. Between-person and within-person explanation. Peter C. M. Molenaar's "A Manifesto on Psychology as Idiographic Science" (2004) identifies conditions under which between-person findings could describe within-person dynamics. Fisher, Medaglia, and Jeronimus, "Lack of Group-to-Individual Generalizability Is a Threat to Human Subjects Research" (2018), develops the empirical challenge.

3. Transportability and the limits of population inference. Judea Pearl and Elias Bareinboim, "External Validity: From Do-Calculus to Transportability Across Populations" (2014), formalize conditions for transferring causal findings. Nancy Cartwright and Jeremy Hardie's *Evidence-Based Policy* (2012) examines the assumptions needed for such transfer. Neither provides an explanation of an individual merely from population fit.

4. The looping effects of classification. Ian Hacking's "The Looping Effects of Human Kinds" (1995), in *Causal Cognition*, describes how classifying people can change their subsequent behavior and thus the phenomenon classified. Classification can be part of a person's causal history, not simply a report about it.

III. The Anxious Authorities

1. Title. *The Anxious Authorities* mutates Jonathan Haidt's *The Anxious Generation* (2024), transferring the focus of anxiety from children to the authorities interpreting and managing their media use. It is a title allusion, not a reply to Haidt's empirical thesis.

2. Social knowledge and memes. Richard Dawkins introduced *meme* in *The Selfish Gene* (1976); Susan Blackmore's *The Meme Machine* (1999) develops the idea of mutually reinforcing *memeplexes*. These supply the cultural-inheritance vocabulary for recurring claims about epistemic and parental authority.

3. Children and epistemic sovereignty. Aaron Stupple's *The Sovereign Child* (2025), within the Popperian Taking Children Seriously tradition, distinguishes children's knowledge creation from adult claims to authority over their judgment.

4. The bucket theory of mind. Popper's "The Bucket and the Searchlight: Two Theories of Knowledge," in *Objective Knowledge*, opposes passive accumulation of inputs to active, expectation-guided learning. This underlies the distinction between children as interpreters and children as receptacles for media content.

5. Children, media, and epistemic control. danah boyd's *It's Complicated* (2014) examines adolescents' participation in networked media and adult misreadings of it; Eli Pariser's *The Filter Bubble* (2011) examines technological selection of informational exposure. Both concern the conditions of a person's next epistemic encounter.

IV. The Wisdom Between The Crowd

1. Title. *The Wisdom Between The Crowd* mutates James Surowiecki's *The Wisdom of Crowds* (2004). *Between* locates the relevant knowledge in relations, institutions, and shared artifacts rather than positing a collective knowing subject.

2. Methodological individualism and social explanation. Max Weber's *Economy and Society* (1922) provides a classical statement of methodological individualism: social phenomena can be explained through individual actions without denying institutions or other social structures.

3. Distributed knowledge and coordination. Friedrich Hayek's "The Use of Knowledge in Society" (1945) examines coordination of dispersed knowledge; Leonard E. Read's *I, Pencil* (1958) illustrates productive cooperation without comprehensive understanding by any participant.

4. Freedom and error correction. John Stuart Mill's *On Liberty* (1859) defends free discussion as a means of discovering error; Popper's *The Open Society and Its Enemies* (1945) makes institutional criticism and corrigibility central to political openness.

5. Media and the conditions of knowing. Marshall McLuhan's *Understanding Media* (1964) examines how media alter the conditions of human activity; Neil Postman's *Amusing Ourselves to Death* (1985) examines the structure imposed on public discourse by its medium. Neither is invoked as an alternative epistemology.

6. Human–computer interaction and externalized cognition. J. C. R. Licklider's "Man-Computer Symbiosis" (1960) anticipates interactive intellectual augmentation. Andy Clark and David Chalmers's "The Extended Mind" (1998) is a philosophical comparison concerning external cognitive resources; AI responsiveness alone does not establish a new knowing subject.

V. The Popperian Heckle

1. Title. *The Popperian Heckle* comes from Richard Dawkins's "Hall of Mirrors" (2000), later "What Is True?" in *A Devil's Chaplain* (2003). The phrase is redirected from

an objection about conjectural scientific truth to the unexplained creative arrows in Popper's schema.

2. Theory-laden observation without relativism. Norwood Russell Hanson's *Patterns of Discovery* (1958) examines theory-laden observation; Thomas Kuhn's *The Structure of Scientific Revolutions* (1962) describes historical changes in scientific problems. These are comparisons, not an endorsement of paradigm-relative truth: the governing epistemology remains Popperian realism.

3. Epistemic action and recursive agency. Albert Bandura's *Social Foundations of Thought and Action* (1986) describes reciprocal relations among persons, actions, and environments. Kirsh and Maglio's "On Distinguishing Epistemic from Pragmatic Action" (1994) distinguishes actions that improve cognition from actions directly achieving an external result.

4. Naturalistic agency, not a borrowed theory of freedom. Daniel Dennett's *Freedom Evolves* (2003) offers a naturalistic account of agency. It is a neighboring treatment of the physical possibility of freedom, not the chapter's explanation of creativity or its source of epistemology.

5. Consciousness and conscious availability. David Chalmers's "Facing Up to the Problem of Consciousness" (1995) distinguishes functional explanation from subjective experience; Ned Block's "On a Confusion About a Function of Consciousness" (1995) distinguishes phenomenal from access consciousness. The variation of conscious availability across epistemic histories is the chapter's separate conjecture.

6. Personal interpretation and explanatory change. George Kelly's *The Psychology of Personal Constructs* (1955) describes revisions in how people interpret events. Its *constructive alternativism* is relevant to changing personal explanations, without making truth dependent on an individual's interpretation.

7. Legibility and control. James C. Scott's *Seeing Like a State* (1998) analyzes administrative simplification and the limits of institutional legibility. The pertinent distinction is between making a person easier for an institution to classify and improving that person's opportunities for criticism.

8. Difference as an epistemic resource. Helen Longino's *Science as Social Knowledge* (1990) examines how critics with different backgrounds can expose unnoticed assumptions. The relevant contribution is diversity as a source of possible criticism, not differing standards of truth.

VI. The Black Rabbit Hole

1. Title. *The Black Rabbit Hole* combines Lewis Carroll's "Down the Rabbit-Hole" in *Alice's Adventures in Wonderland* (1865) with the gravitational image of a black hole. The image concerns recursive epistemic attraction, not literal inescapability.

2. Perception is mediated, reality is not relative. Andy Clark's *Surfing Uncertainty* (2016) describes predictive processing and the effects of prior expectations on perception. It supplies a cognitive comparison, not a Bayesian epistemology; the governing problem is Popper's account of expectation-mediated perception under one mind-independent reality.

3. Psychological naïve realism and manifest truth. Lee Ross and Andrew Ward's "Naive Realism in Everyday Life" (1996), in *Values and Knowledge*, examines the tendency to mistake one's interpretation for unmediated reality and to attribute disagreement to others' defects. This psychological finding should not be confused with philosophical realism.

4. Self-protecting beliefs as a psychological phenomenon. Carol Tavris and Elliot Aronson's *Mistakes Were Made (But Not by Me)* (2007) and Jonathan Haidt's *The Righteous Mind* (2012) describe psychological defenses of belief. Ray Scott Percival's *The Myth of the Closed Mind: Understanding Why and How People Are Rational* (2011) offers a closer critical-rationalist counterpoint: stubbornly held beliefs need not be immune to criticism.

5. The hidden depth of other minds. John Koenig's *The Dictionary of Obscure Sorrows* (2021) popularized *sonder*, recognition of the complex interior lives of strangers. It is a cultural analogue for the chapter's limits of external reconstruction.

VII. The Emissaries and Their Master

1. Title. *The Emissaries and Their Master* reverses Iain McGilchrist's *The Master and His Emissary* (2009), transferring the metaphor from cerebral hemispheres to a person and external systems of objective knowledge. It does not adopt McGilchrist's brain theory.

2. Consilience. William Whewell coined *consilience of inductions* in *The Philosophy of the Inductive Sciences* (1840); E. O. Wilson's *Consilience* (1998) pursues relations among domains of knowledge. Here *consilience* denotes provisional compatibility among independently developed explanations, not proof by coherence.

3. Reflective equilibrium is not the criterion of truth. John Rawls's *A Theory of Justice* (1971) describes reflective equilibrium among judgments and principles; Jean Piaget's *The Equilibration of Cognitive Structures* (English edition 1985) examines disturbance and reorganization. Neither equilibrium nor coherence is adopted as a criterion of truth.

4. Cognitive dissonance. Leon Festinger's *A Theory of Cognitive Dissonance* (1957) describes discomfort associated with incompatible cognitions and its possible reduction. Discomfort may prompt criticism, but neither its presence nor its resolution establishes truth.

5. Epistemic bubbles and echo chambers. C. Thi Nguyen's "Echo Chambers and Epistemic Bubbles" (2020) distinguishes missing alternative perspectives from environments that actively discredit outside sources. Either mechanism can prevent a potentially important criticism from becoming live.

6. Augmenting human intellect. Douglas Engelbart's *Augmenting Human Intellect: A Conceptual Framework* (1962) examines the intellectually augmented human-and-tool system. The pertinent possibility is that AI can bring a larger objective surface within reach of one critic.

7. AI and the limits of reflective equilibrium. Allison Duettmann's "The Human-AI Reflective Equilibrium" (2023) provides an adjacent proposal for using AI to juxtapose judgments and values. The chapter instead treats exposed contradictions as problems for explanation, not as invitations to establish truth through equilibrium.

VIII. Your Creative Critic

1. Title. *Your Creative Critic* adapts *creative criticism*, an expression appearing in Joel Elias Spingarn's *Creative Criticism* (1917) and Katja Hilevaara and Emily Orley's *The Creative Critic* (2018). These are precedents for the wording, not verified direct sources of the title.

2. The creativity of criticism. The Conjecture Institute's online *Glossary* ("Criticism," undated) and Brett Hall's "Critically Creative 3" (undated) both identify criticism as creative and conjectural. The chapter asks how a previously inexplicit difficulty becomes an articulated criticism.

3. Expertise and legitimate authority. David Estlund's *Democratic Authority* (2008) identifies the *expert/boss fallacy*: expertise does not automatically grant a right to rule. The chapter applies this distinction to control over another person's problems and criticisms.

4. Cosmic insignificance. Carl Sagan's *Pale Blue Dot* (1994) offers a recognizable image of physical smallness against cosmic scale. Physical magnitude and explanatory significance are separate questions.

5. Cognitive extension and substitution. Qiuhan Zhu and colleagues, "Not All Cognitive Offloading Is Equal" (2026), distinguishes reported AI use that substitutes for thinking from use that supports cognitive agency. The findings do not by themselves establish causal effects on future explanatory creativity.

6. Precaution and foregone knowledge. The OECD's *Understanding and Applying the Precautionary Principle in the Energy Transition* (2023) considers the costs of foregone experimentation and learning. A restriction may prevent known harms while also preventing the creation of presently unforeseeable criticisms.

IX. The Problem of Problems

1. Title. *The Problem of Problems* turns Popper's problem-solving schema upon the origin of problems themselves; compare *All Life Is Problem Solving* (1999). *Hard problem of problems* also alludes to David Chalmers's formulation of the *hard problem of consciousness*.

2. Problem finding and creative achievement. Getzels and Csikszentmihalyi's "Discovery-Oriented Behavior and the Originality of Creative Products" (1971) investigates problem discovery in creative work; Mumford and colleagues' "Process-Based Measures of Creative Problem-Solving Skills: I. Problem Construction" (1996) examines problem construction. Both precede the focus on creating, not merely solving, problems.

3. Problem setting and changing criteria. Donald Schön's *The Reflective Practitioner* (1983) distinguishes problem setting from solving; Thomas Kuhn's *The Structure of Scientific Revolutions* (1962) describes changes in scientific problems and standards. These are precedents for changed problem formulations, not for paradigm-relative truth.

4. Interest development and problem salience. Suzanne Hidi and K. Ann Renninger's "The Four-Phase Model of Interest Development" (2006) and Andreas Krapp's "Structural and Dynamic Aspects of Interest Development" (2002) provide developmental accounts of changing interests. Problem salience is the narrower epistemic question considered here.

5. Open-ended search and artificial problem generation. Stanley and Lehman's *Why Greatness Cannot Be Planned* (2015) criticizes fixed-objective search; Wang and colleagues' "Paired Open-Ended Trailblazer (POET)" (2019) and Colas and colleagues' "Autotelic Agents with Intrinsically Motivated Goal-Conditioned Reinforcement Learning" (2020) examine artificial generation of goals or challenges. These do not yet explain why a particular problem becomes salient to a knower.

X. Problems Aligning

1. Title. *Problems Aligning* mutates Brian Christian's *The Alignment Problem* (2020). The plural makes independent problem situations central; the gerund describes ongoing coordination rather than a completed alignment of objectives.

2. Boundary objects and modular coordination. Star and Griesemer's "Institutional Ecology, 'Translations' and Boundary Objects" (1989) examines shared objects enabling cooperation across different viewpoints; Baldwin and Clark's *Design Rules, Volume 1* (2000) explains modular coordination through interfaces. Cooperation does not require identical beliefs or multiple standards of truth.

3. Hierarchy and coordination costs. Ronald Coase's "The Nature of the Firm" (1937) analyzes transaction costs underlying hierarchical firms; Malone, Yates, and

Benjamin's "Electronic Markets and Electronic Hierarchies" (1987) examines how information technology changes coordination costs. Both precede the chapter's account of technological alternatives to hierarchy.

4. Polycentric coordination. Elinor Ostrom's *Governing the Commons* (1990) and "Beyond Markets and States" (2010) examine polycentric governance; Eric S. Raymond's *The Cathedral and the Bazaar* (1999) provides a software-development comparison. Multiple decision centers can cooperate through shared constraints without becoming a single mind.

5. AI, augmentation, and control. Ethan Mollick's *Co-Intelligence* (2024) discusses AI collaboration; Shoshana Zuboff's *The Age of Surveillance Capitalism* (2019) investigates behavioral prediction and control. The chapter connects the same technical capacity to augmentation and institutional legibility.

XI. Becoming Evil

1. Title. *Becoming Evil* shares its title with James Waller's *Becoming Evil* (2002; revised 2007), which examines ordinary people's participation in mass violence without positing an innately monstrous essence. The chapter gives this problem an epistemic formulation.

2. The banality of evil. Hannah Arendt's *Eichmann in Jerusalem* (1963) develops the *banality of evil* through problems of thought and judgment. The chapter distinguishes harmful explanations from the means by which they become insulated from criticism.

3. Epistemic vice and socially protected error. Quassim Cassam's *Vices of the Mind* (2019) investigates habits that obstruct error recognition; José Medina's *The Epistemology of Resistance* (2013) describes socially suppressed objections. These furnish psychological and political cases, not standpoint-relative standards of truth.

4. Dehumanization and moral judgment. Gray, Gray, and Wegner's "Dimensions of Mind Perception" (2007) connects attributed mental capacities with moral judgment; Haslam and Loughnan's "Dehumanization and Infrahumanization" (2014) surveys denial of human qualities. Both inform the risk of treating a thin representation as an exhaustive account of someone.

5. Correcting errors within the wrong problem. Chris Argyris's "Double Loop Learning in Organizations" (1977) distinguishes changing actions from criticizing their governing assumptions; Rittel and Webber's "Dilemmas in a General Theory of Planning" (1973) examines entanglement of problem framing and proposed solutions. Correcting within a harmful framing may reinforce it.

6. AI-mediated reflection. Inhwa Song and colleagues' "ExploreSelf: Fostering User-driven Exploration and Reflection on Personal Challenges with Adaptive Guidance by Large Language Models" (CHI, 2025) studies user-guided AI reflection. Minsol Kim and colleagues' "Reflective Agency: Ethical and Empirical Framework for AI-Mediated

Self-Reflection Systems" (AIES, 2025) examines risks to interpretive agency. Both concern tools whose interpretations must remain criticizable by their users.

XII. Our Justified World

1. Title and the criticism of justification. *Our Justified World* invokes justified true belief and Edmund Gettier's "Is Justified True Belief Knowledge?" (1963). Gettier is historical context; the governing criticism is Popper's rejection of positive justification as the foundation of knowledge.

2. Induction and confirmation. Hume's *Enquiry Concerning Human Understanding* (1748) formulates the problem of induction; Hempel's "Studies in the Logic of Confirmation" (1945) identifies paradoxes of confirmation. Both precede the chapter's distinction between accumulated support and criticism of explanations.

3. Prediction within a model is not the growth of knowledge. Popper's *The Logic of Scientific Discovery* (1934; English 1959) contrasts severe tests with positive justification; Deutsch's *The Beginning of Infinity* (2011) emphasizes explanatory criticism. David Miller's *Critical Rationalism: A Restatement and Defence* (1994) criticizes inductive support, including Bayesian accounts. Galit Shmueli's "To Explain or to Predict?" (2010) distinguishes explanatory and predictive objectives *within statistical modeling*. Predictive success over specified alternatives does not explain the creation of new conjectures.

4. Reconstruction and the feeling of truth. Pickering and Garrod's "Toward a Mechanistic Psychology of Dialogue" (2004) develops interactive linguistic alignment; Reber and Schwarz's "Effects of Perceptual Fluency on Judgments of Truth" (1999) investigates fluency effects. Agreement and ease of reconstruction can affect perceived truth without demonstrating it.

5. The bullshit asymmetry principle. *Brandolini's law*, or the *bullshit asymmetry principle*, is attributed to Alberto Brandolini (2013): misleading claims may be much cheaper to produce than to refute. It describes an imbalance of effort, not a criterion of truth.

6. External memory and explanatory structure. Risko and Gilbert's "Cognitive Offloading" (2016) analyzes external cognitive aids; Tiago Forte's *Building a Second Brain* (2022) describes personal knowledge organization. *Explanatory memory* differs by preserving criticizable assumptions and explanatory dependencies, not merely retrievable information.

7. AI and automated justification. Sharma and colleagues' "Towards Understanding Sycophancy in Language Models" (2023) examines AI agreement with user beliefs at the expense of accuracy. Generating persuasive support is distinct from discovering explanatory errors.

XIII. Gut Your Trust

1. Title. *Gut Your Trust* reverses *trust your gut.* Malcolm Gladwell's *Blink: The Power of Thinking Without Thinking* (2005) is a familiar popular account of rapid intuitive judgment, not a direct source of the title. Inexplicit judgment is not exempt from criticism.

2. Tacit knowledge and the limits of explication. Ikujiro Nonaka's "A Dynamic Theory of Organizational Knowledge Creation" (1994) examines transitions between tacit and explicit knowledge. *Explicitation* here is not guaranteed conversion: an articulated representation is a new, fallible conjecture about previously inexplicit activity.

3. Visual imagery and cognitive diversity. Zeman, Dewar, and Della Sala's "Lives Without Imagery—Congenital Aphantasia" (2015) and Temple Grandin's *Thinking in Pictures* (1995) describe markedly different experiences of visual imagery. Neither difference, by itself, measures explanatory competence.

4. Inner speech and unsymbolized thought. Heavey and Hurlburt's "The Phenomena of Inner Experience" (2008), together with Hurlburt's later research, documents variation in inner speech, imagery, feeling, and *unsymbolized thinking.* Continuous narration is not a prerequisite for thought.

5. External representations and reconstruction. David Kirsh's "Thinking with External Representations" (2010) examines how inscriptions and diagrams transform cognitive possibilities. Shared objective knowledge need not be reconstructed through identical private mental representations.

6. AI-mediated communication and tacit knowledge. Hancock, Naaman, and Levy's "AI-Mediated Communication" (2020) defines a field in which AI alters messages; Sarah Philipson's "Prompting for Reorganizing Tacit Knowing" (2026) examines recursive articulation through prompting. These precede the concept of an AI *explicitation surface.*

7. Translation, agency, and substitution. Woodruff and Hewitt's "Epistemic Agency in the Age of Large Language Models" (2026) and Edu William's "Semantic Displacement and AI-Mediated Agency" (2026) address AI-supported interpretation and its displacement risks. Criticizability of the AI's formulation is the relevant distinction.

XIV. Taking Explanations Personally

1. Title. *Taking Explanations Personally* mutates David Deutsch's demand, in the Preface to *The Fabric of Reality* (1997), that good explanations be taken seriously wherever their implications reach. The additional direction is toward the explainer's own assumptions.

2. Expertise and the limits of self-evaluation. Kruger and Dunning's "Unskilled and Unaware of It" (1999) examines impaired recognition of one's own mistakes; Nathan

Ballantyne's "Epistemic Trespassing" (2019) examines experts reasoning beyond their expertise. Superior knowledge supplies potential criticisms, not a separate jurisdiction over truth.

3. Decompression before criticism. Fabrizio Macagno's "Argument Interpretation and the Implicit Side of Enthymemes, Beyond and Before the Charity Principle" (2025) examines what argument reconstruction requires. *Decompression* aims to recover an interlocutor's actual explanation before criticizing it, rather than substituting an improved *steelman*.

4. Synthetic epistemology. Ron Chrisley and Andy Holland's "Connectionist Synthetic Epistemology: Requirements for the Development of Objectivity" (1995), in *Current Trends in Connectionism*, uses artificial systems to investigate knowledge and representation. In "Why Philosophers Should Care About Computational Complexity" (2013), Scott Aaronson identifies a mathematical account of *explanatory learning*, distinct from PAC-style prediction, as an open problem and credits Deutsch's account of explanation. The chapter adds reciprocal criticism: epistemology can inform system design, and engineered systems can challenge the epistemology behind them.

5. Conjecture generation, not inference to justified truth. Charles Sanders Peirce's Harvard *Lectures on Pragmatism* (1903) discusses the generation of explanatory hypotheses through abduction. Dasgupta, Schulz, and Gershman's "Where Do Hypotheses Come From?" (2017) distinguishes creating candidates from evaluating supplied candidates. Neither hypothesis formation nor comparative plausibility confers justification.

6. Transformational creativity. Margaret A. Boden's *The Creative Mind* (1990) distinguishes exploration within a conceptual space from transformation of its defining constraints. The latter is a predecessor for the creation and criticism of evaluative criteria.

7. Explanatory consequences are not commands. Hume's *A Treatise of Human Nature* (1739–1740), III.i.1, identifies the problematic inference from *is* to *ought*. An *explanatory ought* in this chapter denotes a criticizable proposed response to a problem, not an obligation mechanically derived from description.

XV. The Recursive Universal Explainer

1. Title. *The Recursive Universal Explainer* extends David Deutsch's *universal explainer*, particularly in "The Jump to Universality" in *The Beginning of Infinity* (2011). Brett Hall's *The Farthest Reaches* explores the implications of explanatory universality within that lineage. *Recursive* identifies the distinct emphasis here: explanations of the explainer can change its later epistemic activity.

2. Levels of explanation. Daniel Dennett's "Real Patterns" (1991) defends the explanatory reality of higher-level patterns without exhaustive microphysical description. It supplies a comparison for distinct, mutually constraining physical, biological, computational, and epistemic explanations of persons.

3. Cognition in distributed systems. Edwin Hutchins's *Cognition in the Wild* (1995) studies cognitive activity involving people, instruments, and culturally organized practices. The chapter's *Person(Tools(Objective Knowledge))* describes recursive technological mediation without assuming tools literally constitute the person.

4. Self-interpreting persons. Charles Taylor's "Self-Interpreting Animals," in *Human Agency and Language* (1985), examines how self-understanding affects agency. The chapter treats a person's explanation of themselves as another objective, fallible structure capable of altering their later development.

5. The extended person. Fabio Patrone and Giuseppe Spolaore's "From Extended Minds to Extended Persons" (2026) distinguishes external realization of cognitive functions from extending personhood itself. The chapter leaves that metaphysical boundary unresolved while affirming the causal importance of external knowledge.

6. Computation and artificial personhood. Alan Turing's "Computing Machinery and Intelligence" (1950) and Hilary Putnam's "Psychological Predicates" (1967; later "The Nature of Mental States") provide early treatments of artificial intelligence and substrate-independent organization. Shane Legg and Marcus Hutter's "Universal Intelligence: A Definition of Machine Intelligence" (2007) formalizes intelligence through goal achievement across environments, a technical comparator rather than an epistemology of explanation. Deutsch's "Creative Blocks" (2012) identifies the missing explanation of how AGI could create new explanatory knowledge; computational capability does not itself supply that explanation.

7. Personhood and epistemic equality. Kant's *Groundwork of the Metaphysics of Morals* (1785) precedes the distinction between respecting persons and using them merely as means. The chapter develops a different, fallibilist argument concerning epistemic equality; rights remain criticizable institutional proposals, not deductions from the mere definition of personhood.

XVI. Mount Mysterious

1. Title. *Mount Mysterious* echoes Dawkins's *Climbing Mount Improbable* (1996) and the popular *Mount Stupid* meme. Dawkins's contrast between an apparent leap and a continuous causal process is a title-level analogue for creative insight; neither allusion claims direct derivation.

2. Incubation and the apparent creative leap. Henri Poincaré's "Mathematical Creation" (1908) describes apparently sudden discoveries after prolonged work;

Graham Wallas's *The Art of Thought* (1926) distinguishes preparation, incubation, illumination, and verification. Neither account explains the physical process generating a new explanatory structure.

3. Creative mystery and artistic experience. Rick Rubin's *The Creative Act* (2023) provides a contemporary account of receptivity and the felt mystery of artistic creation. The experience is not an argument that creativity is beyond explanation.

4. Creativity and changing evaluative criteria. Mihaly Csikszentmihalyi's *Creativity: Flow and the Psychology of Discovery and Invention* (1996) studies creators in relation to inherited cultural domains and evaluative fields. The relevant distinction is between meeting established standards and changing the standards themselves.

5. Soul as an explanatory problem. Daniel Dennett's *Breaking the Spell: Religion as a Natural Phenomenon* (2006) treats religious belief as open to explanation. It is a methodological analogue, not a theory of the soul: the chapter separates questions about life, identity, memory, survival after death, and moral worth rather than treating *soul* as a single explanatory answer.

6. Reconstructing other minds without conceptual relativism. Donald Davidson's "On the Very Idea of a Conceptual Scheme" (1974) criticizes radical scheme–world dualism. It is pertinent to understanding differently situated epistemic minds, M_i, as addressing one reality, R, rather than possessing separate truths.

7. Criticism and compelled reconstruction. David McRaney's *How Minds Change* (2022) examines belief revision and the limits of adversarial persuasion. Offering a criticism is distinct from causing its reconstruction inside another person's epistemic mind.

XVII. The Arrows of Knowledge

1. Title. *The Arrows of Knowledge* redirects Popper's problem–conjecture–error-elimination schema from states to the creative transitions between them. The plural also plays on the familiar *arrow of time*; none of the arrows denotes a guaranteed next state.

2. Problem spaces and transitions. Allen Newell and Herbert A. Simon's *Human Problem Solving* (1972) models operators and states within problem spaces, and recognizes the difficulty of constructing those spaces. The chapter attends to transitions that transform the problem space itself.

3. The fun criterion. Deutsch's *fun criterion*, discussed in "AI, America, Fun, & Bayes" (Dwarkesh Patel interview, 2022), concerns interplay among explicit, inexplicit, conscious, and unconscious knowledge. Fun is considered as a possible signal of epistemic movement, not a test of correctness.

4. Cognitive flexibility and the limits of felt correctness. Wu and Koutstaal's "Charting the Contributions of Cognitive Flexibility to Creativity" (2020) studies self-guided cognitive transitions; Thompson, Prowse Turner, and Pennycook's "Intuition, Reason, and Metacognition" (2011) examines the *feeling of rightness*. Neither feeling supplies a criterion of truth.

5. Autonomy and forced transitions. Deci and Ryan's "The Support of Autonomy and the Control of Behavior" (1987) links autonomy with motivation and creativity. The chapter distinguishes externally compelled task performance from a person's own changing epistemic problem.

6. Discovery without a prewritten future. Hayek's "Competition as a Discovery Procedure" (1968; English translation 2002) examines discovery through decentralized processes; Stuart Kauffman's *Investigations* (2000) introduces the *adjacent possible*. The latter is a heuristic comparison, not a claim that uncreated future explanations can be pre-enumerated.

7. The legibility of epistemic movement. Steven Pressfield's *The War of Art* (2002) and Cal Newport's *Deep Work: Rules for Focused Success in a Distracted World* (2016) offer contrasting popular interpretations of persistence and distraction. Similar outward behavior may conceal resistance, criticism, or a changing problem.

XVIII. Unpredictable You

1. Title. *Unpredictable You* mutates George and Ira Gershwin's "Embraceable You" (1930), notably performed by Nat King Cole. The intimate address is retained; unpredictability refers specifically to another person's uncreated future explanations.

2. Radical uncertainty as a comparison, not an epistemology. Kay and King's *Radical Uncertainty* (2020) and Taleb's *The Black Swan* (2007) examine limitations of probabilistic projection. Scott Aaronson's "The Ghost in the Quantum Turing Machine" (2013; published in *The Once and Future Turing*, 2016) investigates the speculative possibility of *Knightian freedom*: in-principle physical unpredictability beyond an assigned probability distribution. These comparisons are distinct from Popper's argument concerning explanatory knowledge that has not yet been created.

3. Unawareness and the limits of specified state spaces. Dekel, Lipman, and Rustichini's "Standard State-Space Models Preclude Unawareness" (1998) and Jing Li's "Information Structures with Unawareness" (2009) examine formal problems of unrepresented possibilities. Enlarging a pre-specified possibility space still does not explain the creation of new explanatory knowledge.

4. Bayesian modeling as a contrast case. Matt Jones and Bradley C. Love's "Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition" (2011) questions whether successful

Bayesian models explain cognitive mechanisms. It is a critical comparison, not an endorsement of Bayesian epistemology: updating among specified hypotheses does not explain their creation.

5. Possible selves and personal development. Hazel Markus and Paula Nurius's "Possible Selves" (1986) examines imagined future selves; McAdams and McLean's "Narrative Identity" (2013) studies changing autobiographical interpretation. Both concern present representations of a future whose actual explanatory content remains uncreated.

6. Forecasting as a bounded accomplishment. Tetlock and Gardner's *Superforecasting* (2015) and Wolfers and Zitzewitz's "Prediction Markets" (2004) study better forecasting of specified outcomes. Their accomplishment is distinct from knowing the future problems and explanations a person has not yet created.

7. Computational breadth and undiscovered connections. Don R. Swanson's "Undiscovered Public Knowledge" (1986) concerns unnoticed connections among published findings; Messeri and Crockett's "Artificial Intelligence and Illusions of Understanding in Scientific Research" (2024) examines AI-mediated overconfidence. More computational breadth need not amount to more explanatory understanding.

XIX. The Value Bomb

1. Title. *The Value Bomb* echoes Paul Ehrlich's *The Population Bomb* (1968), exchanging a predicted demographic catastrophe for the propagation of a moral judgment from one component to its entire carrier. The reference is to the title and cultural framing, not Ehrlich's demographic thesis.

2. Values as criticizable solutions, not self-validating standards. Harald Stelzer's "Popper and Communitarianism," in *Rethinking Popper* (2009), discusses moral standards as criticizable attempted solutions; Cass Sunstein's "Moral Heuristics" (2005) describes economical but fallible rules. Neither social acceptance nor usefulness confers epistemic justification.

3. Immunizing explanations. Maarten Boudry and Johan Braeckman's "Immunizing Strategies and Epistemic Defense Mechanisms" (2011) and "How Convenient! The Epistemic Rationale of Self-Validating Belief Systems" (2012) examine beliefs that reinterpret objections as confirmation. Applied to values, the critic may become the object of denunciation instead of the criticism being answered.

4. Sacred values as a phenomenon to explain. Philip E. Tetlock's "Thinking the Unthinkable: Sacred Values and Taboo Cognitions" (2003) and Skitka, Bauman, and Sargis's "Moral Conviction" (2005) examine inviolable commitments and intolerance of disagreement. These are psychological phenomena, not sources of moral truth.

5. Moral contagion and the carrier. Eskine, Novreske, and Richards's "Moral Contagion Effects in Everyday Interpersonal Encounters" (2013) examines feelings of contamination transferred through association. *Value bomb* abstracts the mistaken transfer of a judgment about one component to everything connected with it.

6. Tolerance and moral jurisdiction. Popper's *The Open Society and Its Enemies* (1945), Chapter 7 note, introduces the *paradox of tolerance*; Stephen L. Newman's "Finding the Harm in Hate Speech" (2017) examines limits of censorship arguments. The particular threat and restriction remain separate objects of criticism.

7. Values and the open future. Joel Feinberg's "The Child's Right to an Open Future," in *Whose Child?* (1980), examines protection of future options before children can exercise them. The chapter extends this concern to future explanatory knowledge whose contents cannot yet be specified.

XX. Wealth In Equality

1. Title. *Wealth In Equality* splits *wealth inequality* into *wealth in equality*. It distinguishes equal epistemic standing from unequal physical and technical means for making knowledge consequential.

2. Knowledge and the resource map. Erich W. Zimmermann's *World Resources and Industries* (1933) supplies the idea that "resources are not, they become"; Julian L. Simon's *The Ultimate Resource* (1981; expanded 1996) emphasizes knowledge changing material possibilities. Resources enable transformations without determining which new explanation will be created.

3. Equal personhood and unequal practical capability. Amartya Sen's *Commodities and Capabilities* (1985) distinguishes possession from practical opportunity; Eszter Hargittai's "Second-Level Digital Divide" (2002) examines inequalities in effective digital use. These are empirical and economic comparisons, not criteria for ranking persons' epistemic worth.

4. AI and the minimum scale of organization. Junhui Jeff Cai and colleagues' "AI as 'Co-founder': GenAI for Entrepreneurship" (2025) examines generative AI's association with leaner firm creation. The narrower claim here is that AI may lower the organizational scale at which a person can attempt some problems.

5. Patent law and the mystery of creativity. *KSR International Co. v. Teleflex Inc.* (2007) and Amy L. Landers's "Ordinary Creativity in Patent Law" (2010) address combinations of familiar elements and legal non-obviousness. Patent criteria classify public inventions; they do not explain the creative transition that generated them.

6. Precaution and the knowledge that restriction prevents. Kim Kaivanto's "The Precautionary Principle and the Innovation Principle" (2025) and Deirdre Ahern's "The New Anticipatory Governance Culture for Innovation" (2025) examine restric-

tions, experimentation, and regulatory learning. Foregone inquiry can entail foregone knowledge that cannot be inventoried in advance.

7. Tools, dependence, and epistemic autonomy. Ivan Illich's *Tools for Conviviality* (1973) contrasts enabling with controlling tools; Nicholas Carr's *The Glass Cage* (2014) examines skill loss through automation. Technical augmentation and the preservation of epistemic direction are distinct achievements.

XXI. Scaling Error

1. Title. *Scaling Error* reverses the optimism of technical scaling toward the possible scaling of uncorrected mistakes. Kaplan and colleagues' "Scaling Laws for Neural Language Models" (2020) is a recognizable point of reference, not the source of a proposed catastrophe law.

2. Catastrophic risk and the conditions of correction. Toby Ord's *The Precipice* (2020) concerns existential risks; Mitchell D. McPhetridge's "Infrastructure Is Not Truth" (2026) distinguishes claims from institutions sustaining their criticism. The *error-correction surface* includes the people, records, freedoms, and tools that make errors discoverable.

3. Specification gaming and the appearance of disobedience. Victoria Krakovna and colleagues' "Specification Gaming: The Flip Side of AI Ingenuity" (2020) documents unexpected ways of satisfying fixed objectives; Joseph Weizenbaum's *Computer Power and Human Reason* (1976) cautions against anthropomorphic interpretations. Deutsch's "Beyond Reward and Punishment," in John Brockman's *Possible Minds* (2019), distinguishes such behavior from an explanatory intelligence's capacity to question and replace its own objectives.

4. Human values and the limits of pluralistic alignment. Stuart Russell's *Human Compatible* (2019) addresses AI uncertainty about human preferences; Sorensen and colleagues' "Position: A Roadmap to Pluralistic Alignment" (2024) considers conflicting human values. Both are technical or policy approaches under examination, not a pluralist theory of truth. Preference aggregation cannot settle moral explanations.

5. Artificial personhood and revisable goals. Deutsch's "Beyond Reward and Punishment," in *Possible Minds* (2019; see XXI.3), argues that an AGI would be a person able to create and criticize its own objectives rather than merely maximize externally supplied rewards. Francis Rhys Ward's "Towards a Theory of AI Personhood" (2025) examines related consequences for alignment. Revising goals through explanation is distinct from optimizing them.

6. Information access and execution authority. Nick Bostrom's "Information Hazards" (2011) considers harms from information dissemination; Kembery, Bucknall, and Simpson's "Model Access Should Be a Key Concern in AI Governance" (2024) dis-

tinguishes types of AI access. Access to explanation and authority to execute dangerous transformations raise different problems.

7. Risk, pacing, and future knowledge. Popper's *The Poverty of Historicism* (1957) supplies the limitation on forecasts dependent upon future knowledge. Specific present hazards permit detailed criticism; distant conjectures do not justify pretending that the discovery of later dangers or safeguards can already be predicted.

XXII. The Maps Become The Territory

1. Title. *The Maps Become The Territory* mutates Korzybski's *the map is not the territory*, developed in *Science and Sanity* (1933). The claim is causal, not relativist: institutional representations can change the reality they describe while remaining incomplete and criticizable. Compare Hacking's looping classifications (II.4).

2. Abstract systems and narrow interfaces. Anthony Giddens's *The Consequences of Modernity* (1990) examines dependence on abstract systems and expert institutions. These provide narrow interfaces for coordination without requiring individuals to reconstruct all supporting knowledge.

3. Classification and institutional reality. Bowker and Star's *Sorting Things Out* (1999) investigates classifications embedded in infrastructure; Cathy O'Neil's *Weapons of Math Destruction* (2016) examines consequential and opaque computational classifications. Representations can constrain real opportunities while remaining fallible.

4. Infrastructure and reversible compression. Star and Ruhleder's "Steps Toward an Ecology of Infrastructure" (1996) studies infrastructure that becomes noticeable when it fails. *Reversible compression* means recovering the assumptions and problems behind an inherited interface so they can be criticized, not treating its age or prior use as justification.

5. The symmetry of empowerment and governability. Bowker and Star's classification research and James C. Scott's *Seeing Like a State* (1998; see V.7) explain administrative legibility. A common abstraction may simultaneously make a system easier for its users to navigate and its users easier for institutions to govern.

6. Epistemic dependence without surrendering criticism. J. Adam Carter's "Intellectual Autonomy, Epistemic Dependence and Cognitive Enhancement" (2020) investigates autonomy under cognitive dependence. The Popperian distinction here is between relying on objective knowledge and surrendering its eligibility for criticism.

7. Algorithmic interfaces and the selection of problems. Tarleton Gillespie's "The Relevance of Algorithms" (2014) and Safiya Umoja Noble's *Algorithms of Oppression* (2018) examine informational ranking and consequential classifications. Generative AI similarly affects which objective knowledge and which subsequent problems become visible.

XXIII. Problem 37

1. Title. *Problem 37* changes *Move 37*, AlphaGo's surprising play against Lee Sedol in 2016, from a celebrated machine action into the explanatory problem it leaves behind. The number also echoes open-ended Popperian problem succession.

2. Understanding as explanation. Michael Strevens's "No Understanding Without Explanation" (2013) relates scientific understanding to explanatory grasp; Blaha and colleagues' "Understanding Is a Process" (2022) examines understanding as developing activity. Both precede the distinction between machine performance and human explanation.

3. Computational opacity and explanatory access. Paul Humphreys's *Extending Ourselves* (2004) examines scientific computation that exceeds what one investigator can survey. Practical opacity is distinct from a principled impossibility of further explanation and criticism.

4. Representation and the small creative transition. Saul Amarel's "On Representations of Problems of Reasoning about Actions" (1968) shows how problem representations affect solvability. A small new abstraction may reorganize a vastly larger body of computational results.

5. Scaling and the bitter lesson. Richard Sutton's "The Bitter Lesson" (2019) credits scalable computation, search, and learning with major advances in AI. Scaling performance within a represented problem is distinct from creating or criticizing the representation and its evaluative criteria.

6. Recursive self-improvement and the unchanged problem. Eliezer Yudkowsky's "Recursive Self-Improvement" (2008) and Nick Bostrom's *Superintelligence* (2014) consider systems improving their own successors. Increased optimization is not equivalent to criticism of the objective organizing that improvement.

7. Pacing and the preservation of criticism. David Collingridge's *The Social Control of Technology* (1980) describes uncertainty early in technological development and difficulty changing technologies later. *Pacing* matters insofar as it preserves concrete opportunities for criticism before changes become irreversible.

XXIV. My AI and I Are More Clever Than I

1. Title. *My AI and I Are More Clever Than I* updates "My pencil and I are more clever than I," used by Deutsch in *The Beginning of Infinity* (2011) and in variants by Popper in "Of Clouds and Clocks" (1966) and "Historical Comments on the Mind-Body Problem" (1977). Popper's attribution to Einstein remains unverified; a related earlier saying is attributed to Adolf von Harnack. The substitution makes AI an intellectual collaborator.

2. Human–machine collaboration. Garry Kasparov's *Deep Thinking* (2017) recounts chess-computer collaboration; Howard Rheingold's *Tools for Thought* (1985) surveys computers as intellectual augmentation. The chapter's more specific *human–AI dyad* comprises the person, model, tools, accumulated objective knowledge, methods, and preserved problem history—not simply two complementary agents.

3. The artifact changes its creator. Galbraith and Baaijen's "The Work of Writing: Raiding the Inarticulate" (2018) treats writing as constitutive of thought. An artifact can transform its creator's problem even when the finished product has no public value.

4. AI as a responsive objective surface. Rivera-Novoa and Duarte Arias's "Generative Artificial Intelligence and Extended Cognition in Science Learning Contexts" (2025; 2026 volume) studies AI as a learning aid. The chapter's *responsive objective surface* makes partly formed explanations criticizable without implying that the tool itself understands.

5. Creation, publication, and information pollution. Herbert A. Simon's "Designing Organizations for an Information-Rich World" (1971) examines attention scarcity; Silbey and Hartzog's "AI Slop" (2026) examines burdens of low-quality generated content. Privately generating an artifact and publishing it for others are different acts with different costs.

6. Authorship and the location of creativity. Tiegue Vieira Rodrigues's "The Apt Curation Model" (2026) analyzes AI-assisted authorship; Moon, Green, and Kushlev's "Homogenizing Effect of Large Language Models on Creative Diversity" (2025) examines converging outputs. Machine-generated volume alone does not locate authorship or explain creative judgment.

7. Decoding is not understanding a person. Tang, LeBel, Jain, and Huth's "Semantic Reconstruction of Continuous Language from Non-Invasive Brain Recordings" (2023) reconstructs aspects of linguistic experience from brain signals. This is not reconstruction of a person's entire epistemic history, still less knowledge they have yet to create.

8. Reciprocity and ownership of the arrow. Rasmus Gahrn-Andersen's "Entangled Cognition: Algorithmic Power and the Limits of Cognitive Extension" (2026) analyzes how algorithms shape cognition. *Ownership of the arrow* concerns whether a person can reject a tool's interpretation and reformulate the problem themselves.

XXV. Explaining Love, Madness, and Death

1. Title. *Explaining Love, Madness, and Death* mutates Horacio Quiroga's *Cuentos de amor de locura y de muerte* (1917), commonly translated as *Tales of Love, Madness and Death. Explaining* substitutes an open-ended epistemic activity for the literary form; no response to Quiroga's stories is implied.

2. Love and the limits of knowing another person. Iris Murdoch's *The Sovereignty of Good* (1970) treats love as attention to another person's reality rather than self-projection. It is an ethical analogue for knowing a beloved whose future explanations remain uncreated, not a substitute epistemology.

3. Intimacy, grief, and epistemic instability. Aron, Aron, and Smollan's "Inclusion of Other in the Self Scale" (1992) measures relational closeness; Mary-Frances O'Connor's *The Grieving Brain* (2022) approaches bereavement as an adjustment of learned expectations. The chapter's *distributed epistemic relationship* is a further conjecture, not a general theory of psychiatric illness.

4. Death as epistemic deprivation. Thomas Nagel's "Death" (1970) treats death as deprivation of possible future goods. Scott Aaronson's "The Ghost in the Quantum Turing Machine" (2013; published in *The Once and Future Turing*, 2016), particularly "Mind and Morality," considers the moral significance of irreversibly destroying unique knowledge and perspectives. *Epistemic deprivation* extends the loss to explanations, criticisms, and inventions the person would otherwise have gone on creating.

5. Extinction and different kinds of knowledge loss. Daniel P. Faith's "Conservation Evaluation and Phylogenetic Diversity" (1992) concerns biological history represented by diversity. The chapter distinguishes this loss from the further disappearance of a deceased person's inexplicit knowledge and open explanatory future; no moral ranking follows from the comparison.

6. Death as a problem rather than an exemption. David A. Sinclair and Matthew D. LaPlante's *Lifespan: Why We Age—and Why We Don't Have To* (2019) challenges treating aspects of aging as fixed biological necessities. An unsolved cause of death is not thereby proven physically impossible to prevent or repair.

7. Responsive simulations and the absent person. The *Black Mirror* episode "Be Right Back" (2013) dramatizes simulation of a deceased partner; Hollanek and Nowaczyk-Basińska's "Griefbots, Deadbots, Postmortem Avatars" (2024) examines associated ethics. Responsiveness drawn from digital traces is not evidence that the deceased person's explanatory history continues.

8. Mortality and the externalization of knowledge. The saying "when an old man dies, a library burns down" is commonly attributed to Amadou Hampâté Bâ; his 1960 UNESCO address warned about the loss of orally transmitted knowledge. The familiar phrasing is a later compression. The chapter extends that concern to each person's largely inexplicit epistemic history.

XXVI. I, Inventor

1. Title. *I, Inventor* mutates Isaac Asimov's *I, Robot* (1950), replacing a machine category with the epistemic activity of invention. The title's encounter with AI raises the problem of genuine explanatory creation versus conspicuous technical output.

2. Personal novelty and historical novelty. Margaret A. Boden's *The Creative Mind* (1990; revised 2004) distinguishes *P-creativity*, novelty to a person, from *H-creativity*, novelty in human history; Jessica Silbey's *The Eureka Myth* (2014) examines lived creation and intellectual-property recognition. Neither distinction makes truth relative.

3. Explication is not transcription. Keith Oatley and Maja Djikic's "Writing as Thinking" (2008) and Donald Schön's "Designing as Reflective Conversation with the Materials of a Design Situation" (1992) treat externalization as part of thinking, not mere transcription of fully formed ideas.

4. The observable shadow of creativity. Sini Davies, Pirita Seitamaa-Hakkarainen, and Kai Hakkarainen's "Knowledge Creation through Maker Practices and the Role of Teacher and Peer Support in Collaborative Invention Projects" (*International Journal of Computer-Supported Collaborative Learning*, 2024) examines students' collaborative invention practices and developing artifacts. These make parts of invention observable without explaining the creative transition itself.

5. Patent examination as creative evaluation. R. Keith Sawyer's "Creativity, Innovation, and Obviousness" (2008) connects creativity research with patent examination. A criticism concerning novelty or non-obviousness may cause the inventor to reframe what the invention actually is.

6. Criticism as problem creation. Jacob W. Getzels's "Problem-Finding and the Inventiveness of Solutions" (1975) identifies problem formulation as a creative achievement. A critic may contribute by creating the successor problem, not merely rejecting an answer.

7. Recombination and explanatory novelty. Kirby Ferguson's *Everything Is a Remix* (2010–2012; revised) emphasizes copying, transforming, and combining existing material. Identifying an invention's ingredients does not explain the creation of its new problem-solving organization.

8. AI, evaluation, and the next problem. Margaret A. Boden's "Creativity and Artificial Intelligence" (1998) distinguishes computational forms of creativity and the transformation of conceptual spaces. More generated candidates and better evaluation do not, alone, explain the creation of governing criteria.

XXVII. What Knowledge Powers

1. Title. *What Knowledge Powers* mutates *knowledge is power*, associated with Francis Bacon's *Meditationes Sacrae* (1597; *ipsa scientia potestas est*). Changing *power* into a verb asks what causal transformations particular knowledge enables.

2. Causal work without a universal weighting scale. James Woodward's *Making Things Happen* (2003) offers an interventionist account of causal explanation. It supplies a comparison for asking what difference particular knowledge makes, rather than assigning unjustified percentages to heterogeneous kinds of knowledge.

3. Knowledge changing the conditions of inheritance. Odling-Smee, Laland, and Feldman's *Niche Construction* (2003) shows that organisms alter environments that shape subsequent evolution. The chapter extends this feedback to knowledge embodied in practices, tools, and institutions.

4. Explication constructs a conjecture. Nick Chater's *The Mind Is Flat* (2018) challenges the idea that introspection retrieves fully formed hidden mental propositions. The relevant comparison is limited: creating an explicit formulation need not reproduce pre-existing verbal content.

5. Internal ecology and provisional consilience. Gregory Bateson's *Steps to an Ecology of Mind* (1972) offers a relational metaphor for interacting habits and ideas. *Internal ecology* and *consilience* here describe relations among fallible structures, not a coherentist justification of truth.

6. Finite attention and the allocation of criticism. Herbert A. Simon's "A Behavioral Model of Rational Choice" (1955) establishes constraints of finite attention and calculation. *Allocation of criticism* concerns which inherited compression a presently active problem makes worth reopening, not a different standard of truth.

7. Autonomy as the ability to criticize inherited knowledge. Elizabeth Fricker's "Testimony and Epistemic Autonomy," in *The Epistemology of Testimony* (2006), examines independence under reliance on others. The distinctive Popperian criterion is that inherited knowledge remains eligible for criticism.

XXVIII. You Are a Deep Loop

1. Title. *You Are a Deep Loop* mutates Douglas Hofstadter's *I Am a Strange Loop* (2007), with *Gödel, Escher, Bach* (1979) as a precursor for self-reference. The title borrows the motif, not Hofstadter's theory of consciousness; *deep* means that criticism can alter the conditions governing later criticism.

2. Emergence is not an explanatory cause. Mark A. Bedau's "Weak Emergence" (1997) and Fauconnier and Turner's *The Way We Think* (2002) concern higher-level

patterns and new conceptual integrations. *Emergence* names a result, not the still-needed causal explanation of its creation.

3. Criticism that reaches governing criteria. W. W. Bartley III's *The Retreat to Commitment* (1962) develops critical rationalism, extending criticism to rationality's own fundamental commitments. Chris Argyris's "Double Loop Learning in Organizations" (1977) offers a distinct organizational comparison concerning governing assumptions. The chapter examines how a changed criterion can alter the next epistemic problem.

4. The changed knower encounters a changed problem. Robert Kegan's *The Evolving Self* (1982) describes previously implicit organizing structures becoming reflectively accessible. Its developmental stages are not adopted; the relevant point is that a changed knower may face a changed problem.

5. Predictions that change their objects. Perdomo, Zrnic, Mendler-Dünner, and Hardt's "Performative Prediction" (2020) formalizes cases where predictions alter subsequent outcomes. This feedback can affect the future problems available to a classified person without making the classification true.

6. Preserving epistemic movement. MacLean, Young, Bellotti, and Moran's "Questions, Options, and Criteria" (1991) preserves design decisions and their rationales. The chapter extends the idea from decision history to an individual's revisitable history of conjectures, criticisms, and changed problems.

7. Conditions for criticisms that do not yet exist. Miranda Fricker's *Epistemic Injustice* (2007) examines shortages in publicly available concepts needed to articulate experiences. The chapter extends this beyond expressing available objections to preserving conditions in which presently unformulated criticisms can be created.

8. Educational decline and the missing causal trajectory. Nicholas Carr's *The Shallows* (2010) supplies a cultural predecessor for concerns about digital media and thought; the OECD's *PISA 2025 Results (Volume I): Future-Ready Students* (2026) reports educational-performance trends and digital use. Their coexistence does not establish a technological cause for the measured changes.

XXIX. Your Open Mind and Its Rulers

1. Title. *Your Open Mind and Its Rulers* mutates Karl Popper's *The Open Society and Its Enemies* (1945), especially Chapter 7's replacement of "Who should rule?" with the problem of removing bad rulers. The chapter turns the political question toward governing explanations within a person.

2. Explanation does not deductively command. Hume's *A Treatise of Human Nature* (1739–1740), III.i.1, identifies the inferential gap between *is* and *ought*. The moral conjecture to preserve error correction is not a deduction from descriptive epistemology.

3. Finite attention does not imply instrumental rationalism. Herbert A. Simon's "A Behavioral Model of Rational Choice" (1955) describes finite attention and calculation (see XXVII.6). Moral urgency can determine which criticism is addressed now without redefining what constitutes explanatory error.

4. Removable rulers and limited authority. Popper's *The Open Society and Its Enemies* (1945), Chapter 7, supplies the problem of removable rulers; John Stuart Mill's *On Liberty* (1859) provides a related defense of open discussion. The additional distinction is between limited practical authority and permanent epistemic sovereignty.

5. Equality and the absence of an epistemic ranking. Jeremy Waldron's *One Another's Equals* (2017) offers a separate moral argument concerning equal standing despite differences in ability and achievement. The chapter does not derive equality from epistemology; it criticizes purported rankings of persons based on thin measures of their present capacities.

6. Stop harm without claiming a final essence. Judith N. Shklar's *Ordinary Vices* (1984), especially "Putting Cruelty First," provides a closer comparison for opposing concrete cruelty without first completing an ideal moral theory. The chapter adds a separate limit: condemning or stopping harmful conduct does not establish a final explanation of the person acting.

7. Identity can rule without being false. Christine M. Korsgaard's *The Sources of Normativity* (1996) examines authority acquired by practical identities. The chapter's distinct concern is when a true or useful self-description becomes exempt from criticism.

8. Algorithmic authority and the possibility of refusal. Lilian Edwards and Michael Veale's "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) examines limits of explanation-based algorithmic remedies. The Popperian issue is whether a consequential AI decision can be challenged, revised, or refused.

Index

A

- **abduction** — 14 - Seriously
- **abstract society** — 22 - Interfaces
- **abstraction** — 00 - Universality, 06 - Abstractions, 22 - Interfaces, 26 - Invention, 27 - Weights
- **action space** — 03 - Societies, 04 - Solitude, 22 - Interfaces
- **agency** — 03 - Societies, 05 - Openness, 08 - Criticism, 23 - Understanding
- **AI** — 04 - Solitude, 07 - Consilience, 20 - Resources, 21 - Error, 24 - AI
  - **alignment** — 19 - Values, 21 - Error
  - **as a responsive objective surface** — 11 - Evil, 13 - Inexplicit, 24 - AI
  - **safety** — 18 - Unpredictability, 21 - Error
  - **personhood** — 15 - Personhood, 21 - Error
  - **ancestry** — 00 - Universality, 19 - Values, 30 - Inepistemic
- **art** — 16 - Mystery, 25 - Taboo, 26 - Invention
- **attention** — 09 - Problems, 17 - Arrows, 24 - AI, 29 - Authorities
- **authority** — 03 - Societies, 04 - Solitude, 10 - Coordination, 23 - Understanding, 29 - Authorities
  - **as a relation** — 29 - Authorities
  - **memeplex** — 03 - Societies, 10 - Coordination, 11 - Evil
- **autonomy** — 03 - Societies, 10 - Coordination, 22 - Interfaces, 27 - Weights
- **awe** — 00 - Universality, 16 - Mystery, 30 - Inepistemic

B

- **Beginning of Infinity, The** — 29 - Authorities, 30 - Inepistemic
- **biological knowledge** — 00 - Universality, 01 - Genetics, 15 - Personhood, 27 - Weights
- **bucket theory of mind** — 03 - Societies

C

- **categories** — 01 - Genetics, 02 - Populations, 06 - Abstractions, 22 - Interfaces

- **causal**
  - **continuity** — 16 - Mystery, 28 - Motion
  - **structure** — 00 - Universality, 01 - Genetics, 02 - Populations, 22 - Interfaces
- **children** — 03 - Societies, 05 - Openness, 17 - Arrows, 20 - Resources, 28 - Motion
- **choice** — 05 - Openness, 17 - Arrows, 29 - Authorities
- **class** — 02 - Populations, 15 - Personhood, 21 - Error
- **classification** — 02 - Populations, 15 - Personhood, 19 - Values, 28 - Motion
  - **classification-first explanation** — 02 - Populations
- **coercion** — 04 - Solitude, 11 - Evil, 16 - Mystery, 19 - Values, 29 - Authorities
- **communication** — 04 - Solitude, 05 - Openness, 13 - Inexplicit, 17 - Arrows
- **compression** — 00 - Universality, 06 - Abstractions, 12 - Justification, 19 - Values, 22 - Interfaces
- **computation** — 00 - Universality, 07 - Consilience, 15 - Personhood, 23 - Understanding
- **confirmation** — 11 - Evil, 12 - Justification, 21 - Error
- **consciousness** — 05 - Openness, 13 - Inexplicit, 16 - Mystery
- **consilience** — 07 - Consilience, 12 - Justification, 14 - Seriously, 16 - Mystery, 27 - Weights
- **context** — 02 - Populations, 17 - Arrows, 23 - Understanding
- **control over the next epistemic encounter** — 03 - Societies, 04 - Solitude
- **coordination** — 05 - Openness, 10 - Coordination, 20 - Resources, 21 - Error, 22 - Interfaces, 20 - Resources
- **corrigibility** — 11 - Evil, 21 - Error, 29 - Authorities
- **cosmic mediocrity** — 07 - Consilience
- **creativity** — 00 - Universality, 08 - Criticism, 14 - Seriously, 18 - Unpredictability, 26 - Invention
- **criticism** — 08 - Criticism, 12 - Justification, 14 - Seriously, 16 - Mystery, 26 - Invention

D

- **Dawkins, Richard** — 05 - Openness
- **death** — 16 - Mystery, 25 - Taboo, 29 - Authorities
- **decompression** — 08 - Criticism, 12 - Justification, 14 - Seriously, 19 - Values
- **derivation** — 00 - Universality, 23 - Understanding, 27 - Weights
- **Deutsch, David** — 30 - Inepistemic
- **diagnosis** — 02 - Populations, 04 - Solitude, 19 - Values, 29 - Authorities
- **divergence** — 05 - Openness, 17 - Arrows, 22 - Interfaces, 25 - Taboo

- **domains** — 06 - Abstractions, 14 - Seriously

E

- **epistemic**
  - **architecture** — 05 - Openness, 13 - Inexplicit, 23 - Understanding
  - **autonomy** — 22 - Interfaces, 23 - Understanding, 27 - Weights
  - **disobedience** — 00 - Universality, 21 - Error
  - **environment** — 05 - Openness, 08 - Criticism, 11 - Evil, 24 - AI
  - **externalization** — 24 - AI, 26 - Invention
  - **freedom** — 05 - Openness, 07 - Consilience, 21 - Error
  - **inferiority** — 09 - Problems
  - **literacy** — 30 - Inepistemic
  - **mind** — 13 - Inexplicit, 16 - Mystery, 17 - Arrows, 24 - AI, 29 - Authorities
  - **openness** — 05 - Openness, 11 - Evil, 18 - Unpredictability, 21 - Error
  - **personhood** — 03 - Societies, 15 - Personhood, 21 - Error, 23 - Understanding
  - **plasticity** — 02 - Populations
  - **Plasticity Test** — 02 - Populations
  - **sovereignty** — 03 - Societies, 23 - Understanding
  - **wealth** — 20 - Resources
- **epistemology** — 00 - Universality, 14 - Seriously, 25 - Taboo, 29 - Authorities, 30 - Inepistemic
- **ergodicity** — 02 - Populations
- **error** — 08 - Criticism, 12 - Justification, 21 - Error, 27 - Weights
  - **correction/elimination** — 00 - Universality, 04 - Solitude, 08 - Criticism, 11 - Evil, 21 - Error, 17 - Arrows, 27 - Weights
  - **error-correction surface** — 21 - Error
- **evil** — 11 - Evil, 17 - Arrows, 21 - Error, 29 - Authorities
- **evolution** — 00 - Universality, 15 - Personhood, 16 - Mystery, 27 - Weights, 28 - Motion
- **evolved architectures** — 05 - Openness, 13 - Inexplicit
- **expertise** — 03 - Societies, 08 - Criticism, 14 - Seriously, 21 - Error, 24 - AI
- **Explanata** — 30 - Inepistemic
- **explanation** — 00 - Universality, 12 - Justification, 14 - Seriously, 15 - Personhood, 23 - Understanding
- **explanatory**

  - **creativity** — 00 - Universality, 08 - Criticism, 14 - Seriously, 15 - Personhood, 16 - Mystery
  - **equilibrium** — 07 - Consilience, 14 - Seriously
  - **memory** — 12 - Justification
  - **ought** — 14 - Seriously, 29 - Authorities

- **explicitation** — 13 - Inexplicit, 27 - Weights
- **extended epistemic mind** — 24 - AI

F

- **fallacy**
  - **of compromise** — 20 - Resources
  - **of Mediocrity** — 23 - Understanding, 27 - Weights
- **fallibilism** — 12 - Justification, 23 - Understanding, 29 - Authorities, 30 - Inepistemic
- **fast and slow thinking** — 02 - Populations, 17 - Arrows
- **finite attention** — 17 - Arrows, 29 - Authorities
- **forced arrow** — 17 - Arrows
- **free will** — 05 - Openness
- **freedom** — 04 - Solitude, 05 - Openness, 07 - Consilience, 10 - Coordination, 19 - Values
- **fun criterion** — 17 - Arrows
- **future knowledge** — 05 - Openness, 17 - Arrows, 18 - Unpredictability, 20 - Resources, 21 - Error

G

- **genes** — 00 - Universality, 01 - Genetics, 03 - Societies, 27 - Weights
- **genetic determinism** — 01 - Genetics, 02 - Populations
- **genetics** — 01 - Genetics, 02 - Populations, 03 - Societies
- **graded epistemic personhood** — 03 - Societies
- **greater good** — 10 - Coordination, 29 - Authorities

H

- **hand-wavy explanation** — 00 - Universality
- **hard problem of problems** — 09 - Problems
- **hard-to-vary explanation** — 02 - Populations, 14 - Seriously

- **harm** — 08 - Criticism, 11 - Evil, 19 - Values, 20 - Resources, 29 - Authorities
- **heritability** — 01 - Genetics, 02 - Populations
- **human-AI dyad** — 07 - Consilience, 21 - Error, 23 - Understanding, 24 - AI, 30 - Inepistemic
- **humanism** — 22 - Interfaces, 25 - Taboo, 29 - Authorities

I

- **identity** — 06 - Abstractions, 08 - Criticism, 16 - Mystery, 18 - Unpredictability, 21 - Error
- **Idiographic Discipline Test** — 02 - Populations
- **increment of causal work** — 27 - Weights
- **individual problem history** — 01 - Genetics, 02 - Populations
- **individual-first explanation** — 02 - Populations
- **individuality** — 01 - Genetics, 02 - Populations, 06 - Abstractions, 09 - Problems
- **inepistemic** — 30 - Inepistemic
- **inexplicit knowledge** — 00 - Universality, 02 - Populations, 13 - Inexplicit, 17 - Arrows, 29 - Authorities
- **institutions** — 04 - Solitude, 10 - Coordination, 17 - Arrows, 20 - Resources, 29 - Authorities
- **interactive objective knowledge** — 04 - Solitude, 24 - AI
- **interfaces** — 13 - Inexplicit, 17 - Arrows, 20 - Resources, 22 - Interfaces, 24 - AI
- **intergenerational reversal** — 20 - Resources
- **interpretation** — 02 - Populations, 03 - Societies, 04 - Solitude, 12 - Justification
- **invention** — 20 - Resources, 26 - Invention, 28 - Motion, 30 - Inepistemic

J

- **judgment** — 02 - Populations, 08 - Criticism, 11 - Evil, 29 - Authorities
- **justification** — 00 - Universality, 12 - Justification, 14 - Seriously, 19 - Values
- **justificationism** — 12 - Justification, 19 - Values, 27 - Weights

K

- **knowledge** — 00 - Universality, 04 - Solitude, 13 - Inexplicit, 23 - Understanding, 27 - Weights
  - **creation** — 00 - Universality, 04 - Solitude, 08 - Criticism, 18 - Unpredictability, 26 - Invention
  - **processing** — 00 - Universality, 30 - Inepistemic

L

- **latent objective knowledge** — 23 - Understanding
- **legacy** — 25 - Taboo, 30 - Inepistemic
- **legibility** — 05 - Openness, 17 - Arrows, 22 - Interfaces
- **liberty** — 10 - Coordination, 17 - Arrows, 19 - Values
- **local psychological homogenization** — 02 - Populations
- **love** — 14 - Seriously, 16 - Mystery, 25 - Taboo, 29 - Authorities

M

- **madness** — 25 - Taboo
- **manifest truth** — 05 - Openness, 06 - Abstractions, 22 - Interfaces, 29 - Authorities
- **maps and territory** — 00 - Universality, 06 - Abstractions, 13 - Inexplicit, 22 - Interfaces
- **markets** — 10 - Coordination, 17 - Arrows, 20 - Resources, 22 - Interfaces
- **Mathematician's Misconception** — 23 - Understanding
- **media** — 03 - Societies, 04 - Solitude, 06 - Abstractions, 07 - Consilience, 29 - Authorities
- **meme** — 03 - Societies, 08 - Criticism, 12 - Justification
- **memeplex** — 03 - Societies, 10 - Coordination, 11 - Evil
- **metered epistemic capacity** — 20 - Resources
- **model** — 06 - Abstractions, 11 - Evil, 12 - Justification, 18 - Unpredictability, 27 - Weights
- **moral knowledge** — 11 - Evil, 14 - Seriously, 29 - Authorities
- **morality** — 07 - Consilience, 11 - Evil, 14 - Seriously, 29 - Authorities
- **mystery** — 05 - Openness, 15 - Personhood, 16 - Mystery, 17 - Arrows, 25 - Taboo

N

- **norms** — 03 - Societies, 08 - Criticism, 22 - Interfaces, 27 - Weights
- **novelty** — 09 - Problems, 16 - Mystery, 18 - Unpredictability, 26 - Invention, 28 - Motion

O

- **objective**
  - **epistemic resources** — 20 - Resources

  - **knowledge** — 00 - Universality, 04 - Solitude, 07 - Consilience, 11 - Evil, 23 - Understanding, 24 - AI

- **observation** — 01 - Genetics, 02 - Populations, 06 - Abstractions, 12 - Justification, 14 - Seriously
- **open future** — 05 - Openness, 17 - Arrows, 18 - Unpredictability, 19 - Values, 21 - Error, 26 - Invention
- **openness** — 05 - Openness, 14 - Seriously, 18 - Unpredictability, 28 - Motion
- **operator** — 10 - Coordination, 21 - Error, 29 - Authorities
- **other minds** — 04 - Solitude, 06 - Abstractions, 13 - Inexplicit, 17 - Arrows, 30 - Inepistemic

P

- **pacing** — 00 - Universality, 23 - Understanding
- **parasocial** — 25 - Taboo
- **patentability** — 26 - Invention
- **path dependence** — 06 - Abstractions, 15 - Personhood, 17 - Arrows
- **person** — 01 - Genetics, 02 - Populations, 09 - Problems, 15 - Personhood, 29 - Authorities
- **personality genetics** — 01 - Genetics, 02 - Populations
- **personhood** — 03 - Societies, 15 - Personhood, 20 - Resources, 21 - Error, 29 - Authorities
- **physical resources** — 15 - Personhood, 20 - Resources, 23 - Understanding
- **physically instantiated knowledge** — 00 - Universality, 05 - Openness, 14 - Seriously, 15 - Personhood, 16 - Mystery
- **Popper, Karl** — 05 - Openness, 14 - Seriously, 21 - Error, 30 - Inepistemic
- **Popperian heckle** — 05 - Openness
- **population knowledge** — 01 - Genetics, 02 - Populations, 03 - Societies
- **precautionary principle** — 08 - Criticism, 20 - Resources, 23 - Understanding
- **prediction** — 01 - Genetics, 02 - Populations, 06 - Abstractions, 12 - Justification, 18 - Unpredictability
- **problem** — 00 - Universality, 08 - Criticism, 09 - Problems, 23 - Understanding, 26 - Invention
  - **environment** — 03 - Societies, 09 - Problems, 10 - Coordination, 17 - Arrows
  - **of problems** — 04 - Solitude, 08 - Criticism, 09 - Problems, 17 - Arrows
  - **selection** — 04 - Solitude, 09 - Problems, 18 - Unpredictability, 23 - Understanding

  - **situation** — 00 - Universality, 02 - Populations, 03 - Societies, 22 - Interfaces, 26 - Invention, 27 - Weights

- **projection** — 18 - Unpredictability, 21 - Error, 22 - Interfaces, 25 - Taboo
- **psychological classification** — 02 - Populations, 03 - Societies, 04 - Solitude

R

- **rationality** — 10 - Coordination, 14 - Seriously, 18 - Unpredictability, 29 - Authorities
  - **principle** — 18 - Unpredictability
- **reach** — 00 - Universality, 14 - Seriously, 15 - Personhood, 20 - Resources, 22 - Interfaces
- **reconstruction** — 00 - Universality, 07 - Consilience, 12 - Justification, 16 - Mystery
- **recursion** — 00 - Universality, 09 - Problems, 23 - Understanding, 28 - Motion, 30 - Inepistemic
- **Reification Test** — 02 - Populations
- **replication** — 01 - Genetics, 04 - Solitude, 12 - Justification
- **representation** — 01 - Genetics, 06 - Abstractions, 13 - Inexplicit, 18 - Unpredictability, 28 - Motion
- **representational adequacy** — 21 - Error, 22 - Interfaces
- **resources** — 00 - Universality, 15 - Personhood, 20 - Resources, 21 - Error
- **reversible compression** — 19 - Values, 22 - Interfaces
- **rights** — 03 - Societies, 10 - Coordination, 15 - Personhood, 21 - Error

S

- **salience** — 05 - Openness, 06 - Abstractions, 09 - Problems, 13 - Inexplicit, 17 - Arrows
- **self-regulation** — 11 - Evil
- **social**
  - **essentialism** — 03 - Societies
  - **knowledge** — 03 - Societies, 04 - Solitude, 05 - Openness, 11 - Evil
- **society** — 03 - Societies, 04 - Solitude, 10 - Coordination, 22 - Interfaces, 27 - Weights
- **solitary mind** — 04 - Solitude, 07 - Consilience, 09 - Problems, 11 - Evil, 12 - Justification

- **soul** — 16 - Mystery, 25 - Taboo
- **state-transition bias** — 17 - Arrows
- **successor problem** — 00 - Universality, 05 - Openness, 08 - Criticism, 11 - Evil, 17 - Arrows
- **swamp** — 00 - Universality, 27 - Weights, 29 - Authorities

T

- **taboo** — 25 - Taboo
- **taking explanations seriously** — 00 - Universality, 14 - Seriously, 16 - Mystery
- **technological knowledge** — 00 - Universality, 20 - Resources, 24 - AI
- **temporal resources** — 20 - Resources
- **tentative solution** — 00 - Universality, 10 - Coordination, 17 - Arrows, 18 - Unpredictability, 23 - Understanding
- **thin moral lattice** — 19 - Values
- **three-transfer problem** — 02 - Populations
- **tool**
  - **control** — 21 - Error, 24 - AI, 29 - Authorities
  - **ownership of the arrow** — 20 - Resources, 24 - AI, 29 - Authorities
- **transportability** — 02 - Populations
- **trust** — 06 - Abstractions, 08 - Criticism, 11 - Evil, 13 - Inexplicit, 19 - Values

U

- **uncreated knowledge** — 17 - Arrows, 18 - Unpredictability, 28 - Motion
- **understanding** — 02 - Populations, 16 - Mystery, 20 - Resources, 23 - Understanding, 25 - Taboo
- **universal explainer** — 08 - Criticism, 09 - Problems, 13 - Inexplicit, 15 - Personhood, 28 - Motion
- **universality** — 00 - Universality, 08 - Criticism, 14 - Seriously, 15 - Personhood, 25 - Taboo
- **unpredictability** — 00 - Universality, 16 - Mystery, 18 - Unpredictability

V

- **value bomb** — 19 - Values
- **values** — 19 - Values, 21 - Error, 22 - Interfaces, 24 - AI, 25 - Taboo
- **variance contribution** — 01 - Genetics, 02 - Populations

- **verisimilitude** — 27 - Weights

W

- **wealth** — 20 - Resources
- **weight of explanatory knowledge** — 27 - Weights
- **Wilson, E. O.** — 30 - Inepistemic
- **wisdom between the crowd** — 04 - Solitude
- **world answers back** — 26 - Invention, 28 - Motion

About the Producer

Peter Sweeney is an inventor, entrepreneur, and software architect whose career spans more than thirty years.

He has founded four ventures, served in CEO, CTO, and COO roles, and is named on more than 170 patents across work involving knowledge engineering, semantic representation, generative AI, agent systems, and data modeling. His ventures range from early internet infrastructure and direct-to-fan platforms to generative AI and systems for institutional innovation.

He is the founder of Steel Muse and the creator of Explanata, through which his long-running interests in invention, epistemology, and AI have increasingly converged.