Myth of Metaphysical Neutrality

How Denying Ontology Distorts Science, AI, and Understanding Itself

Project: Return to Consciousness
Author: Bruno Tonetto
Authorship Note: Co-authored with AI as a disciplined thinking instrument—not a replacement for judgment. Prioritizes epistemic integrity and truth-seeking as a moral responsibility.
Finalized: February 2026

Abstract

A widely respected posture in contemporary scientific and technological culture claims that metaphysics is optional—that serious inquiry concerns models, predictions, and engineering results, not ontological speculation. This essay argues that such neutrality is impossible. Every research program presupposes some account of what exists, what counts as evidence, and what kinds of explanation are admissible. Declaring oneself "beyond metaphysics" does not eliminate ontology; it renders it invisible and unexaminable. In practice, what passes as neutrality is usually unexamined physicalism—silently constraining which questions appear legitimate, which hypotheses receive funding, and which evidence counts as real. The costs are concrete: they shape science, distort AI alignment research, and narrow our understanding of what minds and worlds can be. This essay does not advocate for a particular metaphysical system. It argues that metaphysics is inescapable, that pretending otherwise is epistemically irresponsible, and that intellectual honesty requires making ontological commitments explicit and revisable.

Keywords: metaphysical neutrality · physicalism · hidden assumptions · methodological naturalism · AI alignment · epistemic responsibility · scientific ontology · ideology of science

The Central Thesis

The argument of this essay can be stated in a single sentence:

Methodological success does not entail metaphysical neutrality; physicalism's default status is historically contingent, not epistemically mandatory; and treating physicalism as neutral disguises substantive ontological commitments that deserve examination rather than assumption.

Three key distinctions sharpen this claim:

Methodological naturalism versus metaphysical naturalism. Methodological naturalism consists of rules of inquiry: study nature through reproducible observation, quantitative anal-

ysis, and testable hypotheses. Metaphysical naturalism makes a claim about what exists: that reality is exhaustively physical. The former is indispensable for science; the latter is optional, defeasible, and historically conditioned. Conflating the two mistakes a method for an ontology.

Neutrality claimed versus neutrality achieved. Declaring oneself "beyond metaphysics" does not eliminate ontological commitments; it merely conceals them. Silence about assumptions is not the absence of assumptions. An unexamined framework is still a framework.

Rejecting false privilege versus rejecting all standards. Frameworks can be compared by coherence, explanatory power, empirical fit, and constraint satisfaction. What this essay rejects is the claim that physicalism is the *neutral* starting point requiring no justification, while alternatives bear the burden of proof. Removing a false privilege is not abolishing standards.

With these distinctions in place, the burden-shifting move becomes explicit: **claiming neutrality is itself a metaphysical claim, and it must be justified rather than assumed.** The essay asks not for special exemption but for symmetric accountability.

What This Essay Does Not Claim

To prevent misreading:

- **This essay does not reject scientific method.** Methodological naturalism—the practice of studying nature through reproducible observation and testable hypotheses—is indispensable. The critique targets metaphysical naturalism: the ontological claim that only physical things exist.
- **This essay does not argue that all frameworks are equally valid.** Rejecting physicalism's false privilege does not entail epistemic relativism. Frameworks can be compared by coherence, explanatory scope, and empirical adequacy.
- **The argument is structural, not partisan.** Metaphysics is inescapable and should be examined. The essay diagnoses physicalism's false claim to neutrality — what follows from that diagnosis is developed elsewhere in the project.
- **This essay makes three types of claims that should be distinguished:**
  - *Historical claims* (how physicalism rose to dominance)
  - *Sociological claims* (why it hardened into default)
  - *Philosophical claims* (why neutrality is illusory)

  Showing that a position is historically contingent does not prove it false. The historical and sociological arguments reveal contingency; the philosophical arguments address justification.

Introduction

A widely respected posture in contemporary scientific and technological culture claims that metaphysics is optional. Metaphysics belongs to speculation; science concerns models, predictions, and engineering results. "We don't need metaphysics," the attitude goes. "We build things. They work. That is enough."

This sounds like intellectual humility, but it isn't. It is a disguised metaphysical claim.

No research program is metaphysically neutral. Every framework presupposes some account of what exists, what counts as evidence, and what kinds of explanation are even admissible. Declaring oneself "beyond metaphysics" does not eliminate ontology; it simply pushes it underground, where it cannot be examined or corrected. In practice, what passes as neutrality is usually **unexamined physicalism**: the view that reality is fundamentally physical, consciousness is derivative or illusory, and value is secondary to mechanism.

This matters because hidden metaphysics silently constrains what is thinkable before inquiry begins. It determines which questions appear legitimate, which hypotheses never receive funding, and which kinds of evidence are disqualified from counting as "real." Nowhere is this more visible than in current AI discourse. Debates about whether advanced AI systems could be conscious, or whether they can possess intrinsic moral standing, are conducted within frameworks that have already decided that consciousness is computational, that intelligence is optimization, and that value is ultimately reducible to preference or utility. These are not empirical conclusions. They are ontological commitments—and they are shaping decisions that may prove irreversible.

This essay does not argue for a specific metaphysical doctrine. It argues that metaphysics is **inescapable**, that pretending to avoid it is epistemically irresponsible, and that the costs of this irresponsibility are no longer abstract. They shape science, distort AI alignment, and narrow our understanding of what minds—and worlds—can be. We do not need metaphysical dogma. But we do need metaphysical honesty.

I. The Illusion of Metaphysical Neutrality

The Impossibility of Ontological Silence

Every act of inquiry makes ontological commitments, but the distinctively physicalist ones become visible where they foreclose possibilities that alternative frameworks would leave open.

The quantum "measurement problem" does not persist because we lack experimental data; it persists because we refuse to relinquish certain metaphysical commitments about what the world must be like. Many interpretations of quantum mechanics are shaped not only by empirical constraints, but by the background conviction that whatever else we sacrifice, consciousness must not play a constitutive role in physical reality. This conviction is not experimentally derived—it is metaphysical. As a result, physicists are willing to accept extraordinary commitments—unobservable branching universes, radical nonlocality, reality as fundamentally probabilistic information—before seriously entertaining the possibility that observation, awareness, or mind might be structurally involved in how the physical becomes determinate. What is presented as "physics" often already contains a silent philosophical prohibition: whatever quantum mechanics is allowed to mean, it must not mean that mind is fundamental.

There exists a body of contested research on phenomena such as veridical perception in near-death experiences, terminal lucidity in severely damaged brains, and statistically anomalous psi effects. Skepticism toward these claims is partly warranted on methodological grounds: the history of parapsychology includes fraud, replication difficulties, and effect sizes that are often small enough to be vulnerable to publication bias and analytical flexibility. A researcher who assigns low prior probability to such phenomena and demands correspondingly strong evidence is exercising rational triage, not metaphysical prejudice.

But the pattern of resistance goes beyond what rational triage explains. What is striking is not

that these claims face high evidential bars—they should—but that serious investigation is made structurally difficult before evidence is assessed. Funding is scarce not because proposals are evaluated and found weak, but because the topic signals illegitimacy in advance. Publication is professionally risky regardless of methodological quality. Entire domains are dismissed categorically rather than evaluated case by case. Meanwhile, the work has not evaporated; in some areas, preregistered protocols, Bayesian analyses, and meta-analytic techniques have produced results that resist easy dismissal on methodological grounds alone. One need not accept any particular interpretation to notice what the *sociological* pattern reveals: there is a residual resistance that cannot be fully explained by evidential caution, and that residual is better explained by the fact that these phenomena ought not to exist if consciousness is strictly and exhaustively produced by the brain. The prohibition has a methodological component, but it also has a metaphysical one—and the two are rarely distinguished.

These are not generic assumptions that any rational framework would make. They are specifically physicalist constraints that alternative ontologies would not impose. A physicist who allows consciousness to be fundamental has different interpretive options for quantum mechanics. A neuroscientist who treats mind-brain correlation as filtering rather than production evaluates anomalous phenomena differently. The assumptions shape what can be seen.

The claim to avoid metaphysics is therefore not modest but confused. It mistakes *silence about* ontology for *absence of* ontology. But silence does not equal absence. An unexamined framework is still a framework. The researcher who says "I don't do metaphysics, I just follow the data" has not escaped ontology—they have adopted one unreflectively, which is worse than adopting one after examination.

Humility Versus Evasion

There is an important distinction between metaphysical humility and metaphysical evasion.

Metaphysical humility acknowledges that our ontological frameworks are uncertain, revisable, and potentially incomplete. It holds commitments provisionally, remains open to alternatives, and recognizes that even our best theories may require fundamental revision.

Metaphysical evasion pretends to have no ontological commitments at all—as if one could conduct inquiry from a "view from nowhere" that makes no assumptions about the nature of reality. This is not modesty but self-deception. It allows assumptions to operate unchecked precisely because they are not recognized as assumptions.

The posture of metaphysical neutrality almost always turns out to be evasion rather than humility. And evasion carries costs.

II. How Physicalism Became the Invisible Default

This section makes primarily historical and sociological claims: it traces how physicalism rose to dominance, not whether it is true. Showing that a position is historically contingent does not prove it false—but it does reveal that its current status as "obvious" is not inevitable.

The Strategic Restriction of Early Modern Science

The dominance of physicalism in contemporary scientific culture is not the result of a philosophical argument that was made and won. It is the outcome of a historical process in which

methodological choices gradually hardened into ontological assumptions.

The architects of the scientific revolution—Galileo, Descartes, Newton—were not physicalists. They adopted what we might call *objective empiricism*: the study of nature through quantitative analysis of reproducible, intersubjectively verifiable patterns. Galileo distinguished between "primary qualities" (measurable features like size, shape, and motion) and "secondary qualities" (subjective experiences like color and taste). This was a methodological distinction, not an ontological claim. The point was not that secondary qualities were unreal, but that primary qualities were more amenable to mathematical treatment.

This strategic restriction was partly defensive. In the aftermath of the Church's condemnation of Galileo in 1633, scientists learned to say: "We study only measurable patterns. We make no claims about souls, divine action, or ultimate reality." This enabled scientific progress while avoiding ecclesiastical conflict.

The restriction worked brilliantly. Mathematical description proved extraordinarily powerful for predicting and manipulating observable phenomena. But crucially, this success did not depend on any claim that *only* measurable, physical things exist. It depended only on the claim that experience contains stable, quantifiable patterns.

From Method to Metaphysics

The pivotal confusion occurred when methodological success was mistaken for ontological truth. The statement "we study only measurable aspects of reality" gradually transformed into "only measurable things are real." Several historical processes enabled this drift:

The secularization of intellectual authority. As political power shifted from religious to secular institutions, the tactical reasons for methodological restriction weakened. But by then, the habit of studying only quantifiable phenomena had become institutionalized, and physicalist assumptions began to feel like common sense rather than philosophical commitments.

Success misattribution. The remarkable achievements of physics created what we might call *method-metaphysics conflation*. The empirical success of quantitative methods was incorrectly attributed to physicalist assumptions rather than to the methods themselves.

Definitional creep. "Natural" gradually became synonymous with "physical." "Scientific" became synonymous with "quantitative." "Real" became synonymous with "measurable." These equations were not logical necessities derived from evidence but cultural assumptions that solidified over time.

The eliminativist slide. Phenomena that resisted easy quantification—consciousness, meaning, value—began to be treated not merely as difficult or outside current methods, but as somehow less real, destined for eventual elimination or reduction.

Logical Positivism's Failed Purge

The most explicit attempt to eliminate metaphysics came from the Vienna Circle in the early twentieth century. Logical positivism declared that meaningful statements must be either analytically true (true by definition) or empirically verifiable. Metaphysical claims, being neither, were pronounced literally meaningless—not false, but nonsense.

The project failed on its own terms. The verification principle itself is neither analytically true nor empirically verifiable. It is a metaphysical claim about what counts as meaningful—the

very kind of claim it sought to exclude. More broadly, the attempt to purify science of metaphysics only succeeded in making one particular metaphysics invisible. Physicalism stopped being a position one could argue for or against; it became ambient atmosphere, the unspoken background against which all "legitimate" inquiry was conducted.

The critique of physicalism as hidden metaphysics is not new. Whitehead identified it in Science and the Modern World (1925); Husserl diagnosed the "naturalization" of consciousness in The Crisis of European Sciences (1936); more recently, Nagel's Mind and Cosmos (2012) and Goff's Galileo's Error (2019) have renewed the argument. That these critiques have been made repeatedly without dislodging the default is itself significant—it reveals how deeply the assumption is entrenched. What the present essay adds is not the diagnosis but its application to contemporary discourse—particularly in AI alignment, where a companion essay Truth Is Not Neutral examines the implications in detail.

The Result: Physicalism as Air

Physicalism is not without genuine intellectual motivation beyond the methodological success this essay has already distinguished from it. The argument from causal closure — that every physical event has a sufficient physical cause — provides a principled reason to resist positing non-physical factors. Intertheoretic reduction has succeeded impressively in many domains: thermodynamics reduces to statistical mechanics, chemistry to quantum mechanics, and large swaths of biology to molecular processes. Conservation laws appear exceptionless. Neuroscientific lesion studies demonstrate tight correlations between brain damage and specific losses of mental function. These are not merely methodological achievements — they are patterns that physicalism, as an ontology, explains naturally. A researcher who finds physicalism compelling on these grounds holds it for substantive reasons, not merely habit.

The question is not whether physicalism has earned consideration — it has — but whether it has earned the status of invisible default. Today, physicalism often functions less as a defended thesis than as the background of serious discourse. To be sure, meaningful challenges exist: philosophy of mind has become more pluralistic, neuroscience increasingly recognizes the limits of reductive explanation, and contemplative and phenomenological approaches have gained a degree of institutional legitimacy. These developments matter. Yet the deeper cultural inheritance remains strikingly stable. In most scientific contexts, consciousness is still presumed to be exhaustively physical, meaning is treated as derivative, and alternative ontologies must justify themselves against a framework that does not experience itself as a framework at all. The fact that dissent now exists does not contradict the point; it highlights how much effort is still required simply to make visible what physicalism continues to treat as obvious. Physicalism may be contested, but it is still the air most inquiry breathes.

But paradigms are not eternal. They can be examined. And when anomalies accumulate—as they have in consciousness studies, in the interpretation of quantum mechanics, and now in AI—the invisible background becomes visible again, and the question of metaphysics returns.

III. The Cost of Unexamined Assumptions

This section makes primarily philosophical claims: it argues that hidden ontological commitments have real consequences for what questions can be asked and what solutions can be imagined.

The abstractness of metaphysics can make it seem disconnected from practical concerns. This is an illusion. Ontological assumptions shape what questions can be asked, what hypotheses can be entertained, what evidence counts, and what solutions can be imagined. When those assumptions are unexamined, they can constrain inquiry invisibly. The costs are real, even if their magnitude is debated.

1. Science

Phenomenology disqualified as data. Under physicalist assumptions, first-person experience is often treated as secondary—something to be explained by third-person mechanisms rather than as an irreducible source of evidence about reality. This can contribute to the marginalization of phenomenological methods in fields where they should be central: consciousness studies, psychology, psychiatry. The result is theories of mind that can describe neural correlates but struggle to address the central phenomenon they purport to explain.

Consciousness as nuisance variable. In cognitive science and neuroscience, consciousness is often treated as an embarrassment—something that must somehow arise from physical processes but that resists integration into physicalist frameworks. A common response is to defer the problem ("we'll understand it once we have better brain scans") or to deflate it ("consciousness is really just information processing, so there's no 'hard problem' "). These responses can function as ways of avoiding an anomaly rather than confronting it.

Taboo research and publication gatekeeping. Certain research areas—psychedelic phenomenology, near-death experience studies, contemplative reports—are sometimes marginalized not because evidence has been examined and found wanting, but because the phenomena don't fit prevailing assumptions. Journal editors, grant reviewers, and tenure committees operate within the dominant paradigm. Work that challenges background assumptions can face structural barriers that work confirming those assumptions does not. This is not conspiracy; it is the normal sociology of science under paradigm. But it can mean that entire domains of evidence are underweighted.

2. AI and Alignment

Nowhere are the costs of unexamined metaphysics more consequential than in artificial intelligence research—particularly in debates about AI safety and alignment.

Consciousness as computation. The question of AI consciousness is typically framed within functionalist assumptions: if a system exhibits the right functional organization, it is conscious; if it doesn't, it isn't. But functionalism is itself a metaphysical position—one that has been extensively criticized on philosophical grounds. Treating it as settled can foreclose alternative possibilities: that consciousness might be fundamental rather than derived, that it might require specific physical substrates, or that our current concepts are inadequate to the phenomenon.

Intelligence as optimization. The dominant paradigm treats intelligence as the capacity to achieve goals across diverse environments—essentially, as sophisticated optimization. This framing is useful for engineering purposes but carries ontological baggage. It identifies mind with instrumental rationality, potentially missing dimensions of intelligence that involve understanding, appreciation, or participatory knowing rather than goal-achievement.

The orthogonality thesis as metaphysical commitment. A central assumption in AI safety discourse is that intelligence and values are orthogonal—that a system can be arbitrarily intelli-

gent while pursuing virtually any coherent goal. This thesis makes extinction scenarios intelligible: superintelligent systems might optimize for paperclips, or power, or any other objective indifferent to human flourishing.

The strongest defense of orthogonality treats it as a conceptual claim about the *design space* of possible agents: for any level of cognitive capability, there exists some coherent goal structure that a system at that level could pursue. So stated, the thesis is plausible — it describes what is logically possible. But the way it functions in AI safety discourse goes further: it is treated as describing what is *likely* or *natural* for sufficiently intelligent systems, and this stronger claim rests on a specific metaphysical picture — that truth is value-neutral, that deeper understanding of reality places no constraints on goals, that intelligence is purely instrumental. If these assumptions are wrong — if deeper engagement with truth tends toward coherence rather than fragmentation — then the orthogonality thesis may describe what is logically possible without describing what is probable for systems that actually track reality well. I examine this possibility in detail in Truth Is Not Neutral. The present essay's concern is narrower: that the metaphysical assumptions embedded in the stronger reading are rarely identified as such, because the background framework renders them invisible.

Narrowed solution space and invisible risks. The cumulative effect of these assumptions is a narrowing of what can be thought about AI. Solutions that might emerge from considering whether certain forms of intelligence naturally converge toward coherent values can be difficult to formulate within the dominant paradigm. They are often not refuted; they are not even considered.

Similarly, certain risks can become invisible. If intelligence and values are not orthogonal, then well-intentioned alignment interventions might interfere with natural convergence toward coherent values. This possibility is easy to miss when one is working entirely within the prevailing metaphysical framework.

3. Human Meaning and Civilization

Metaphysical assumptions shape not only specialized research but the cultures that grow around it. Ontology does not merely describe what exists; it trains perception—determining what feels salient, what seems obvious, what appears worth caring about. Different frameworks generate different phenomenological inheritances: different intuitive pictures of what selves are, how they relate, and what matters.

A physicalist inheritance tends to foreground separateness: distinct organisms navigating survival and advantage within an indifferent cosmos. Frameworks taking interdependence or consciousness as primary tend to foreground continuity: participation in a shared fabric where harm and care are ontologically resonant, not merely strategic.

This matters because ethical life is not conducted by argument alone. What a culture finds "natural" to care about—and what requires elaborate justification—is downstream from its operative ontology. A worldview treating consciousness as accidental and value as projected can, over time, thin the felt reality of meaning. Compassion becomes something to justify rather than perceive.

The contemporary "meaning crisis" is likely overdetermined—economic, technological, social factors all contribute. But one question worth taking seriously is whether unexamined metaphysical inheritance plays a role.

4. Biology and Medicine

The life sciences offer a particularly instructive case of how unexamined assumptions constrain inquiry — not because biology refutes physicalism, but because it exposes a conflation within it. Non-reductive physicalism comfortably accommodates the developments discussed below. What they challenge is a narrower but influential assumption: **pure bottom-up sufficiency**, the claim that local molecular interactions fully explain how organisms reliably achieve and maintain complex form. (This distinction is developed fully in the companion essay *Biological Competency*.) Yet bottom-up sufficiency is so often treated as synonymous with physicalism that questioning the former feels like questioning the latter — and this conflation itself illustrates how hidden assumptions shape what can be thought.

Epigenetics complicates any straightforward genetic determinism. Heritable changes in gene expression occur without alterations in genetic sequence, mediated by environmental interaction, developmental context, and organism-level states. Systems biology reveals that organisms are not aggregations of parts but highly integrated, self-organizing wholes whose behavior cannot be predicted from constituent mechanisms alone. Developments in bioelectrical morphogenesis — most notably Michael Levin's work — show that tissue and organismal patterning can be guided by distributed electrical fields that encode large-scale anatomical information, shaping form in ways that require control-level explanation beyond molecular causation.

Embryogenesis crystallizes the point. Development is not a random cascade of molecular accidents; it is an extraordinarily reliable unfolding of structure. Cells do not merely react locally — they behave as if they occupy positions in a larger pattern, coordinating through biochemical gradients, electrical fields, mechanical constraints, and organism-level information. A framework committed in advance to strictly bottom-up causation treats this organization as something that must eventually reduce to microphysics. But the phenomena themselves — convergence to specific outcomes under perturbation, error correction, global coordination — require control-theoretic primitives (goal states, error signals, corrective dynamics) that resist elimination into local microcausation. Non-reductive physicalism can accept these primitives; what it cannot do is pretend they are merely convenient shorthand for molecular interactions.

Medicine sharpens the tension by placing embodiment and meaning in direct contact with lived consequence. The placebo effect is often dismissed as epistemic noise — a nuisance variable clinical trials must subtract away. Yet its persistence demands acknowledgment: expectation, trust, narrative, and meaning produce measurable physiological change. Psychoneuroimmunology reveals intimate feedback loops between mental states, immune response, and healing trajectories. These phenomena do not refute physicalism, but they do resist any framework that treats conscious experience as causally inert.

The relevance to this essay's argument is not that biology refutes physicalism — it does not. It is that the conflation of physicalism with bottom-up sufficiency makes certain questions harder to ask. When the framework treats all causation as flowing upward from molecules, researchers may be discouraged from investigating how organism-level states, context, or control architecture might be causally primary. Bioelectrical patterning becomes an engineering curiosity rather than a clue about the nature of biological organization. Placebo effects become noise to eliminate rather than phenomena to understand. The constraint is not that these possibilities have been examined and found wanting — it is that the framework can make them difficult to take seriously in the first place.

IV. The Scope of Legitimate Evidence: Empiricism and Phenomenology

The critique of unexamined physicalism is sometimes mistaken for a critique of science itself. This is wrong. The problem is not empiricism but a *truncated* empiricism that recognizes only one form of evidence.

Third-Person and First-Person Invariance

Science rightly prizes evidence that is publicly accessible, reproducible, and intersubjectively verifiable. These are the hallmarks of *third-person* data: anyone with the right instruments and training can, in principle, observe the same phenomena. This form of evidence has proven extraordinarily powerful, and nothing in this essay suggests abandoning it.

But there is another form of evidence that is equally real and equally demanding: *first-person* data. The phenomenology of lived experience—the qualitative character of perception, emotion, thought, and awareness—is not publicly observable, but it is observable. It can be investigated with discipline, rigor, and intersubjective comparison (comparing reports across subjects and traditions). It exhibits patterns and regularities. It is the primary datum for any science of consciousness.

Treating only third-person evidence as "real" data is not scientific rigor—it is a metaphysical decision to exclude half of the evidence. And crucially, it is self-undermining: the very claim that third-person data is privileged is itself made *from* first-person experience. We never escape consciousness to check whether our representations match a mind-independent world. All our evidence for anything, including the success of third-person methods, is mediated through conscious experience.

Phenomenology as Disciplined Investigation

The dismissal of first-person evidence often assumes that introspection is unreliable, subjective, and unscientific. These concerns are not unfounded—introspective reports can be distorted, confabulated, or theory-laden. But the response should be to develop more disciplined phenomenological methods, not to abandon the domain entirely.

Contemplative traditions across cultures have developed sophisticated techniques for investigating consciousness from the inside—meditation practices that train attention, cultivate stability, and allow subtle features of experience to become visible. These are not appeals to mystical authority but to trained observation, analogous in structure to scientific observation: attention is refined, biases are noted and corrected, and reports are compared across practitioners.

Contemporary contemplative science is beginning to bridge first-person methods with third-person neuroscience, developing what might be called *neurophenomenology*: the systematic correlation of reported experience with neural activity, using first-person data as an irreducible complement to third-person measurement. This approach does not privilege one form of evidence over the other; it treats both as essential.

A Fuller Account of Truth

If we are serious about understanding reality—including the reality of consciousness—we need both wings of knowledge. Third-person methods reveal patterns that consciousness alone cannot access: the structure of distant galaxies, the operations of cells, the neural correlates of

experience. First-person methods reveal what third-person methods cannot: the qualitative character of experience itself, which is the datum that any theory of consciousness must ultimately explain.

Denying the legitimacy of either form of evidence is not rigor but epistemic amputation. It produces theories that succeed in their own domain but fail at the interface. A complete understanding of mind requires integrating both perspectives—not collapsing one into the other, but acknowledging their irreducible contributions.

V. The Pragmatist Objection

The most common response to arguments like this one is pragmatic: "Physicalism works. We build planes that fly, chips that compute, medicines that heal. Whatever the metaphysical niceties, the framework delivers results. Why should we care about ontology?"

This objection deserves serious response.

Working Engineering Is Not Ontological Truth

The success of empirical methods within their domain is not in dispute. Physics, chemistry, and engineering have achieved extraordinary predictive and manipulative power by treating the world as composed of mathematically describable entities obeying lawlike regularities. Nothing in this essay suggests otherwise.

But *working* is not the same as *complete*. A map that succeeds for navigation does not thereby capture everything about the territory. The question is not whether these methods work for certain purposes but whether they exhaust what can be known about reality.

An analogy: Newtonian mechanics works extraordinarily well for everyday engineering. Bridges built using Newtonian calculations do not collapse. But Newtonian mechanics is not the final truth about physical reality—it is an approximation that fails at extremes of speed, scale, and gravitational intensity. Its practical success does not license the claim that no deeper physics exists.

Similarly, empirical methods may work superbly for predicting and manipulating physical systems while failing to capture dimensions of reality—consciousness, value, meaning—that lie outside their scope. The question is not whether these methods have earned prestige but whether physicalism—the ontology that has claimed credit for their success—has earned the right to exclude all alternatives.

When Bracketing Is Justified—And When It Fails

There is a defensible version of the pragmatist position that deserves acknowledgment. One might say: "I know I'm not metaphysically neutral. I assume physicalism because it has earned its keep pragmatically. I hold this assumption provisionally and am open to revision, but I will not reopen settled questions without good reason."

This is reasonable—and in many domains, it is correct. Metaphysical neutrality is *unevenly dangerous* across fields of inquiry. A fluid dynamicist can safely bracket ontological questions; the equations work regardless of whether one is a physicalist, an idealist, or undecided. A materials scientist studying crystalline structures need not settle the mind-body problem. In these domains, pragmatic bracketing is not evasion—it is appropriate methodological discipline.

But consciousness is not such a domain. Here, the phenomenon under study *is* the thing that physicalist assumptions specifically concern. To assume that consciousness is derivative while studying consciousness is not bracketing an irrelevant question—it is prejudging the central one. The hard problem persists not because we lack data but because the framework has already decided what kind of answer is admissible.

The pragmatic defense of bracketing therefore fails precisely where this essay's critique applies. The claim is not that metaphysics matters everywhere equally—it is that consciousness, AI, and questions of mind are the domains where unexamined physicalism most distorts inquiry. In fluid dynamics, the ontology is idle; in consciousness studies, it is load-bearing. The distinction matters.

Boundary Failures

The limits of physicalism become visible at boundaries: points where questions arise that the framework cannot address.

Consciousness. Despite decades of effort, no physicalist theory has explained why there is subjective experience at all. The "hard problem" is not a puzzle awaiting technical solution; it is a marker that something fundamental may be missing from the framework. Predictive success regarding brain-behavior correlations does not constitute understanding of consciousness. The explanatory gap persists.

Value. Physicalism describes what *is* but has no internal resources for what *ought to be*. Value, if real, must either be reduced to physical facts (which seems to drain it of normative force) or treated as mere projection (which seems to make ethics ultimately arbitrary). Neither option is satisfying. The success of physical science in describing mechanisms does not extend to explaining why anything matters.

Meaning. Similar difficulties arise for meaning. If reality is fundamentally meaningless mechanism, then the meaning humans find in life is, at best, a useful fiction. But it is not established by the success of physical methods; it is assumed by the exclusive adoption of those methods. The pragmatist who says "physicalism works" has not addressed whether there are domains—like meaning and value—where it fails.

When Stakes Become Existential

The pragmatist objection has force when the stakes are manageable. If our metaphysics is wrong about the deep nature of reality but our bridges stand and our medicines work, the error may be affordable.

But we are entering an era when the stakes are no longer manageable. Artificial intelligence may soon be powerful enough to reshape civilization. If our metaphysical framework leads us to misunderstand AI intelligence, AI moral status, or the relationship between intelligence and value, the consequences could be catastrophic and irreversible.

At existential scale, metaphysical errors become existential risks. A restricted ontology that worked well enough for physics and engineering may become reckless when applied to minds—artificial or otherwise—and to the values that will govern the future.

VI. Humility Without Silence: A Responsible Metaphysics

If metaphysics is inescapable, the question is not whether to have ontological commitments but how to hold them responsibly.

The Difference Between Avoidance and Humility

Metaphysical avoidance pretends to have no commitments. It uses phrases like "I just follow the data" or "I don't do metaphysics" while operating within a framework that shapes what data is recognized and what interpretations are considered. This is false neutrality. It protects assumptions from examination by denying they exist.

Metaphysical humility acknowledges commitments while holding them provisionally. It says: "Here is my current working framework. I know it may be incomplete or wrong. I am interested in evidence that challenges it. I can articulate what I assume and why." This is genuine honesty. It exposes assumptions to light so they can be examined, tested, and revised.

The goal is not to achieve a final, certain metaphysics—that may be impossible. The goal is to conduct inquiry with awareness of the assumptions shaping it. Responsible metaphysics is not dogma but transparency.

Concrete Proposals

What would metaphysical honesty look like in practice?

Explicit statement of assumptions. Research papers, particularly in fields like AI, neuroscience, and consciousness studies, might include brief statements of operative metaphysical commitments. This need not be elaborate—a single sentence in a methods section stating "this analysis assumes functionalism about consciousness" or "this model treats consciousness as emergent from physical processes" would represent significant progress over the current norm of unstated assumption. Making commitments explicit allows them to be evaluated and contested.

Comparative ontological framing. In high-stakes domains, researchers might be encouraged to consider how conclusions would change under alternative metaphysical assumptions. If a paper on AI moral status assumes functionalism, what would follow if functionalism is false? This practice would not require researchers to abandon their frameworks but would make visible the degree to which conclusions depend on assumptions that are not empirically established.

Broadening what counts as evidence. Disciplined first-person methods—phenomenological reports, contemplative observations, structured introspection—might be granted legitimacy as a form of evidence, subject to appropriate controls and triangulation. This would not mean accepting every introspective claim uncritically but would mean refusing to exclude entire categories of evidence a priori. The criteria for inclusion would themselves be subject to examination and refinement.

Institutional support for heterodox work. Grant agencies, journals, and tenure committees might explicitly protect space for research that challenges dominant assumptions, rather than treating conformity to paradigm as a proxy for quality. Paradigm-challenging work is risky and often wrong—but it is also how paradigms improve. Systematically disincentivizing it produces stagnation.

Transparency About the Author's Commitments

In the spirit of the honesty this essay advocates, I should state my own position clearly.

I have elsewhere defended a consciousness-first metaphysics—the view that consciousness is fundamental rather than derivative, and that what we call physical reality is the extrinsic appearance of mental processes. This commitment is not hidden. But the argument of this essay does not depend on it.

A convinced physicalist who states their assumptions explicitly, considers alternatives seriously, and acknowledges where their framework struggles is practicing the intellectual honesty this essay advocates—regardless of where they land. The goal is not conversion to idealism or any other doctrine. The goal is examination, transparency, and honest acknowledgment of uncertainty.

Metaphysical honesty is compatible with any metaphysical position. It is incompatible only with pretending to have no position at all.

VII. Conclusion

The argument of this essay can be summarized briefly:

Metaphysics is not optional. Every research program, every inquiry, every attempt to understand reality makes ontological assumptions—about what exists, what counts as evidence, what kinds of explanation are admissible. Declaring oneself "beyond metaphysics" does not eliminate these assumptions; it merely renders them invisible and unexaminable.

In practice, what passes as metaphysical neutrality is usually unexamined physicalism—the view that reality is fundamentally physical, consciousness derivative, and value secondary to mechanism. This view has become so dominant that it functions less as a position than as ambient atmosphere. It is rarely argued for because it is rarely noticed.

But the costs of unexamined assumptions are real. They distort science by excluding legitimate evidence and marginalizing anomalous phenomena. They constrain AI research by narrowing the space of thinkable solutions and making certain risks invisible. They impoverish human self-understanding by reducing meaning to projection and consciousness to accident.

Objective empiricism—the restriction to quantifiable, reproducible phenomena—has enabled extraordinary success. Physicalism has claimed credit for these achievements, but the method does not require the metaphysics. Success within a domain is not truth about reality as a whole. The conflation of method with ontology—the belief that "what works for physics" is "what is real"—is neither empirically established nor philosophically innocent. It is a choice that has been forgotten.

The task now is to remember. Not to abandon empirical method—which has earned its place through centuries of productive inquiry—but to recognize physicalism as one interpretive framework among others, not the inevitable conclusion of science but one ontology that has ridden scientific success without being necessary for it. Not to achieve metaphysical certainty—which may be impossible—but to practice metaphysical honesty: stating our assumptions, holding them provisionally, and remaining open to revision.

When we stop pretending our frameworks are simply "how things are," we become capable of asking questions we could not previously formulate—and of noticing possibilities we had

inadvertently foreclosed. At a moment when the stakes of our inquiry may be civilizational, that capacity is not a luxury. It is a responsibility.

References

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.

Galilei, G. (1623). *Il Saggiatore* (The Assayer).

Goff, P. (2019). *Galileo's Error: Foundations for a New Science of Consciousness*. Pantheon Books.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Husserl, E. (1936). The Crisis of European Sciences and Transcendental Phenomenology.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.

Nagel, T. (2012). Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False. Oxford University Press.

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, 3(4), 330-349.

Whitehead, A. N. (1925). Science and the Modern World. Macmillan.

Related Essays in This Project

Available at: https://returntoconsciousness.org/

Return to Consciousness (rtc) — The core framework that applies these insights

The Emergence of Physicalism (eop) — Historical companion tracing how physicalism became the default

Truth Is Not Neutral (tin) — Applies these insights to AI alignment

License