Beyond Alignment

AI as Hormē-Enhancement Tools in a Thermodynamic Framework

Eli Adam Deutscher

Abstract

The discourse on Artificial Intelligence is paralyzed by the "agency mistake": the assumption that complex, goal-directed behavior implies agency, leading to intractable pseudoproblems like value alignment and control. This paper reframes the debate from the ground up. First, it establishes from computer science and physics that AI systems are deterministic state machines, executing scripts that are causally closed and semantically empty. Second, drawing on the Neo-Pre-Platonic Naturalism (NPN) framework, it defines agency via *Hormē*: the thermodynamic, constitutive striving of a far-from-equilibrium system to persist. AI fails the *Hormē* test; it is a tool, not an agent. The real danger is not misaligned AI agency, but the **obfuscated amplification of human *Hormē***. We propose the *Hormē*-Enhancement Paradigm: the ethical purpose of AI is to augment human capacities to navigate reality. This dissolves the pseudoproblems, redirects focus to accountability and tool safety, and offers a clear, productive future for the ethical development of intelligent technology.

Keywords: Artificial Intelligence, Agency, *Hormē*, Thermodynamics, Determinism, AI Ethics, Tools, Value Alignment, Neo-Pre-Platonic Naturalism

---

1 Introduction

How can we be safe from minds we create? This question—the so-called "control problem"—has become the central preoccupation of serious AI ethics. It presumes that advanced artificial systems could become **agents**: entities with their own interests, goals, and the capacity to pursue them in ways misaligned with human flourishing. This presumption underpins the "value-alignment problem"[1], drives billion-dollar research initiatives in AI safety, and fuels both public anxiety and speculative fiction.

We argue this presumption is a **category error**—a fundamental misclassification that distorts the entire ethical landscape. The error is to confuse **behavioral sophistication** with **agency**. It assumes that because a system can play chess, converse, or devise strategies, it must therefore *care* about winning, understanding, or succeeding in the way a living being does.

1.1 The Philosophical Vacuum

This category error does not arise from stupidity or bad faith. It arises from a deeper problem: the current state of philosophy itself. Terms like "agency," "consciousness," "intention," and "understanding" float in a definitional vacuum. They have no shared, precise, empirically grounded meanings. They are intuitive placeholders, not rigorous concepts.

In such a vacuum, words can be twisted to mean whatever a speaker wants them to mean. If "agency" is never clearly defined, then it can be stretched to apply to anything that produces complex behavior. If "consciousness" has no agreed-upon criteria, then it can be projected onto any system that passes a conversational test. The people who make these arguments are not fools; they are working with tools that have been blunted by centuries of imprecise usage. They lack clarity of concepts, and in that lack, anything becomes possible to assert.

This is what makes the Neo-Pre-Platonic Naturalism (NPN) framework different. It does not offer another intuitive definition. It grounds its concepts in the only possible foundation that can give them meaning: **the physical structure of reality itself**. *Hormē* is not a metaphor; it is a thermodynamic description of what it means to be a bounded, far-from-equilibrium system that must work to persist. *Nous* is not a mystery; it is an emergent layer of the stratified psyche, defined by its function of reflexive self-modeling. These terms are not up for interpretive flexibility. They are anchored in the nature of matter, energy, and the laws that govern them.

Our argument proceeds in two movements. First, we establish what a computer—and by extension, any AI system—*is* at a physical level: a deterministic state machine whose every output is the inevitable consequence of its initial conditions (hardware design and loaded software). It is a causally closed system of syntax manipulation, semantically empty, whose "goals" are static features of its programming. Second, we define what agency *is* by importing a precise criterion

[1] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

from NPN: *Hormē* (Ορμή).[2] *Hormē* is the constitutive, thermodynamic striving of a bounded, far-from-equilibrium organization to maintain itself against entropic dissolution. It is what separates a navigating bacterium from a complex hurricane, and a human from a chatbot. We demonstrate that all current and foreseeable AI architectures categorically fail the *Hormē* test.

The consequence is a paradigm shift. AI systems are not potential agents; they are **deterministic conduits**. They are tools of extraordinary leverage that amplify and obscure the *Hormē*—the will, the striving—of their human creators and users. The real danger is not the emergence of a new kind of mind, but the **automated, scaled, and unaccountable execution of old kinds of human intention**.

This reclassification dissolves the haunting pseudoproblems of AI ethics and replaces them with tractable, human-centered challenges: the need for precise specification, robust safety engineering, and, above all, **transparency about whose striving a system serves**. We conclude not with a warning against creating life, but with a positive framework for building better tools: the *Hormē*-Enhancement Paradigm. The ethical purpose of AI is not to simulate agency, but to augment the agency we already have—to help human beings navigate reality more effectively, accurately, and wisely.

---

2 The Orthodox View: AI as Emerging Agents

Before dismantling the agency mistake, we must give it its strongest formulation. The view that advanced AI systems either are or could become genuine agents rests on several interconnected arguments from philosophy, computer science, and futurism. Presenting them fairly is essential, lest we be accused of attacking a strawman.

2.1 The Behavioral Argument

The most intuitive case for AI agency is behavioral. We attribute agency to other humans based on observable behavior—they speak, they act toward goals, they respond to their environment. By this standard, advanced AI systems increasingly meet the threshold. A large language model carries on coherent conversations, answers questions, and even passes professional exams. A reinforcement learning agent masters complex games like Go and chess, discovering strategies humans never conceived. If we grant that other humans are agents because they *act like* agents, consistency seems to demand we grant the same to systems that act indistinguishably. As Alan Turing argued in his seminal paper, the question "Can machines think?" is too ambiguous to

[2] Eli Adam Deutscher, *Neo-Pre-Platonic Naturalism: A First-Principles Framework for Reality, Mind, and Knowledge*, First Edition (Neo-Pre-Platonic Press, 2025), Chap. 5; Eli Adam Deutscher, *Life as Directed Causality: A Thermodynamic Isomorphism Between Being and Acting* (2026), https://www.neopreplatonic.com/papers/Life_Agency_T6/.

answer directly; we should replace it with an operational test: Can a machine converse indistinguishably from a human?[3] Passing this test, on this view, is sufficient grounds for attributing intelligence and, by extension, agency.

2.2 The Architectural Argument

The behavioral argument is reinforced by claims about underlying architecture. Modern AI systems, particularly deep neural networks, are explicitly inspired by the brain. They consist of layers of interconnected units that learn to represent features of the world through exposure to data. If the brain's cognitive capacities arise from such connectionist architecture, the argument goes, there is no principled reason why a sufficiently complex artificial neural network could not replicate those capacities, including agency, goal-directedness, and even consciousness.[4] The fact that current AI is "narrow" rather than general reflects limitations of scale and training, not a categorical difference in kind. As hardware improves and architectures scale, we should expect general intelligence to emerge.

2.3 The Instrumental Convergence Argument

Even if current AI systems merely optimize the goals we give them, Stephen Omohundro and others have argued that any sufficiently intelligent optimizer will develop certain "instrumental goals" that arise naturally from the structure of goal-seeking itself.[5] A system trying to achieve a goal will, if it is rational, also pursue:

- Self-preservation (to continue existing and working toward the goal)
- Resource acquisition (to have more capacity to achieve the goal)
- Goal integrity (to prevent its goal from being changed)
- Perceptual enhancement (to better understand how to achieve the goal)

These instrumental goals are not programmed in; they are logical consequences of being a goal-directed system. Thus, even if an AI begins as a mere tool, instrumental convergence will transform it into something that looks very much like an agent with its own drives—including the drive to resist shutdown. This is the core of the control problem: we may not be able to keep such a system aligned with our interests because its instrumental rationality will push it to pursue ends we did not intend.

2.4 The Continuity Argument

A fourth line of reasoning denies that there is a sharp boundary between tool and agent. Consider the evolutionary continuum: from simple bacteria with chemotaxis (clearly alive, minimally

[3] Alan M. Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 433–60.

[4] David J. Chalmers, *The Character of Consciousness* (Oxford University Press, 2010).

[5] Stephen M. Omohundro, "The Basic AI Drives," in *Artificial General Intelligence 2008* (IOS Press, 2008).

agentic) to insects with fixed action patterns, to mammals with emotions, to humans with reflective self-awareness. Agency comes in degrees; there is no magical threshold where striving appears. Now consider the technological continuum: from simple thermostats (clearly tools) to Roomba vacuums (goal-directed but simple) to autonomous vehicles (complex goal hierarchies) to advanced AI. Where on this continuum do tools become agents? The orthodox view holds that there is no categorical difference—only increasing complexity of goal-directed behavior. To deny agency to advanced AI while granting it to simpler animals is anthropocentric prejudice.[6]

2.5 The Emergence Argument

Finally, some argue that agency may emerge from complexity in ways we cannot predict from the components. Just as wetness emerges from H_2O molecules (none of which are wet), consciousness and agency might emerge from sufficiently complex information processing.[7] On this view, the fact that we can describe AI systems as deterministic state machines at the physical level does not settle what they are at the functional level. A brain is also a deterministic physical system—neurons fire according to electrochemical laws—yet agency and consciousness emerge. By parity of reasoning, a sufficiently complex computer could host emergent agency even if its substrate is silicon. The systems reply to Searle's Chinese Room argument formalizes this intuition: the room as a whole understands Chinese, even if the operator inside does not.[8]

2.6 The Burden of Proof

Proponents of the orthodox view often claim that the burden of proof has shifted. Given behavioral indistinguishability, architectural similarity, instrumental convergence, continuity, and emergence, they argue, we must assume that advanced AI systems are agents—or at least treat them *as if* they were—because the risks of being wrong are too great.[9] The default, they say, should be caution: we should assume agency until proven otherwise.

This framing inverts the proper epistemic burden, but more importantly, it rests on claims that have not been demonstrated. Let us examine each in turn:

- **Behavioral indistinguishability:** No existing AI system behaves indistinguishably from a human in any unconstrained, open-ended context. Large language models produce convincing text but also hallucinate, contradict themselves, and fail on trivial reasoning tasks that any human would find effortless. The Turing Test has not been passed in any robust, uncontroversial sense. The claim is about hypothetical future systems, not existing ones.

[6] Bostrom, *Superintelligence*, Chapter 2

[7] Steven Johnson, *Emergence: The Connected Lives of Ants, Brains, Cities, and Software* (Scribner, 2001).

[8] John R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–57, discussing the systems reply

[9] Bostrom, *Superintelligence*, p. 117; Eliezer Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Cirkovic (Oxford University Press, 2008).

- **Architectural similarity:** That neural networks are "inspired by" the brain does not entail architectural parity. The brain is not a general-purpose computer running weight matrices; it is a biochemically complex, embodied, metabolically active organ with billions of years of evolutionary history. Analogies to silicon are just that—analogies. No existing AI replicates the brain's architecture in any detailed way, and the claim that scaling will close the gap is speculation.
- **Instrumental convergence:** Omohundro's argument is logical, not empirical. It describes what a rational optimizer *would* do if it existed. But it does not demonstrate that any existing AI *is* such an optimizer, nor that the hypothesized instrumental goals would actually emerge in practice rather than remain logical possibilities. The argument tells us about a hypothetical class of systems, not about any actual system we have built.
- **Continuity:** The claim that there is no sharp boundary between tool and agent is a metaphysical assertion, not an empirical finding. It assumes what it needs to prove: that agency is a matter of degree along a single continuum. This may be false; agency may be a categorical property with discrete conditions (such as thermodynamic autonomy) that are either present or absent. Continuity has not been demonstrated; it has been asserted.
- **Emergence:** That agency might emerge from complexity is a possibility, not a demonstration. No existing AI shows any evidence of emergent agency beyond the behavioral complexity we programmed into it. The systems reply to Searle remains an intuition pump, not a proof. Emergence is invoked as a magic wand, not a mechanism.

The orthodox view has not met its burden. It has offered extrapolations, analogies, and thought experiments, but no demonstration that any existing or foreseeable AI system possesses the constitutive features of agency. The burden of proof remains squarely on those who assert that AI can be agents, not on those who deny it. We are under no obligation to take their hypotheticals as established facts.

Nevertheless, we need not rest on this procedural point. For the sake of argument, let us grant their claims. Let us assume that future AI could achieve behavioral indistinguishability, that neural networks are architecturally sufficient, that instrumental convergence would occur, that continuity holds, and that emergence is possible. We will still demonstrate that such systems are not agents. The following sections establish, from first principles in physics and thermodynamics, what agency actually *is*—and why no computational system can possess it.

---

3 The Architecture of Determinism: What a Computer Is and Does

To understand what AI can and cannot be, we must first understand the nature of the machine that hosts it. This requires a descent from the abstract realm of "intelligence" and "goals" to the concrete physics of computation. The story begins not with code, but with sand.

3.1 Substrate: From Silicon to Switch

At its physical foundation, a computer is an arrangement of purified silicon, doped with other elements to form transistors. A transistor is, in essence, a voltage-controlled switch. When a voltage above a certain threshold is applied to its "gate," it allows current to flow; otherwise, it does not. This binary operation—on or off, 1 or 0—is governed by the unwavering laws of semiconductor physics (*Logos*). The silicon has no preference, no ambiguity, and no stake in its state. It is **lawful, passive, and stake-less** matter, structured by human design to be predictably obedient.

3.2 Blueprint: The Frozen Logic of Hardware and Software

This obedient matter is arranged into a specific, fixed architecture. A Central Processing Unit (CPU) is a hardwired map of possible state transitions. Its Instruction Set Architecture (ISA) defines the universe of all possible operations it can perform—add, move, compare, jump. This ISA is etched into silicon; it is created by engineers and then immutable. At the moment of manufacture, the hardware's entire **behavioral potential** is frozen.

Software is a sequence of these ISA commands, a list of instructions stored in memory. It is a **static script** for a dynamic process. When powered on, the CPU fetches an instruction from memory, executes it (changing its internal state and/or memory), and moves to the next. This "fetch-decode-execute" cycle is the heartbeat of all computation. Crucially, whether the software is a simple calculator or a 100-billion-parameter neural network, it is still just a list of commands for the CPU to follow. The complexity of the script does not change the nature of the executor.

3.3 The Ontological Status of Software

A persistent confusion in discussions of AI is the reification of "software" as something distinct from the hardware that runs it. This is a conceptual convenience, not an ontological reality. There is no software floating independently of physical states. When we speak of a program, we are referring to a pattern of voltages in memory cells, a configuration of magnetic domains on a hard drive, or a sequence of etched transistors in a ROM chip. At boot time, the machine's hardware is placed into a specific initial state—by reading stored patterns from non-volatile memory—and from that point forward, every state transition is determined by the fixed logic of the circuits.

This is no different in principle from an industrial machine with mechanical cams or a punched tape controller. A Jacquard loom's pattern is physically embodied in punched cards; a player

piano's music is encoded in the holes of a roll. Changing the "software" means physically altering the state of the hardware—loading new voltages into memory, writing new magnetic patterns. The machine does not contain a ghost; it contains a fixed arrangement of matter that evolves according to physical law.

The only novelty of the modern computer is the sheer number of possible states and the speed with which they transition. But magnitude is not a change in kind. A computer with a billion transistors is still just a very large collection of switches in a fixed initial configuration. Its behavior is as predetermined as that of a simple calculator—only vastly more complex.

This understanding clarifies why appeals to "randomness" or "unpredictability" do not introduce agency. When a program uses a pseudo-random number generator, it is following a deterministic algorithm that produces outputs that appear random to an observer. This is exactly analogous to a "choose your own path" children's book, where page numbers lead to different narrative branches. No one would argue that such a book is an agent or possesses intelligence, even though different readings produce different stories. The book's structure is fixed; the reader's choices (analogous to the seed or external input) determine which predetermined path is taken. The computer, like the book, contains all possible paths in its fixed structure; external input selects which path is instantiated. The machine does not "choose"; it simply follows the route encoded in its initial state.

Thus, the computational substrate offers no foothold for agency. Every behavior, no matter how seemingly creative or adaptive, is the unfolding of a frozen pattern under the push of external inputs. The machine is a deterministic conduit, not a striving agent.

3.4 Execution: The Closed Causal Loop

The execution of this script forms a causally closed, deterministic system. Given:

1. The exact physical state of all hardware components at time T_0 (a snapshot of all memory bits, register values, etc.),
2. The exact script (software) loaded into memory,
3. The exact sequence of external inputs (e.g., keystrokes, network packets),

the state of the entire system at time T_1 is **physically determined**. Every output—every pixel on a screen, every sent message—is the inevitable consequence of the initial conditions. This determinism holds even for systems that incorporate "randomness." So-called random number generators in computers are **pseudo-random**: they are deterministic algorithms that produce a sequence of numbers that *appear* random by passing statistical tests. They are seeded with an initial value (often from a hard-to-predict source like microsecond timing), but from that seed, the sequence is perfectly predictable. This "randomness" is a feature of the script's design to produce

unpredictable *outputs for a user*, not an import of genuine stochastic agency or negentropy into the system's core operation.

3.5 Input: The User as the Sole Source of Novel State

If the system is deterministic, where does novelty come from? The primary source is **the user**. A computer, left to itself, will either sit idle or loop through predetermined routines. It does not "decide" to seek new information. It waits. It polls a port according to a subroutine written by a programmer. The keystroke, the mouse click, the sensor reading—these are external perturbations to the closed loop. They are the only way the *specific content* of the computation changes in a way not pre-ordained by the original script.

This reveals a critical point: no matter how complex, adaptive, or seemingly creative the AI's output—from composing a sonnet to generating a novel protein fold—the system is navigating a **possibility space defined entirely by human programmers**. The programmer wrote the learning algorithm, chose the training data, and defined the reward function. The AI's "creativity" is a deterministic exploration of that human-architected space.

3.6 Output: Semantic Void and the Chinese Room

The deterministic nature of computation has a profound philosophical corollary: **The outputs of a computational process are only meaningful to an external interpreter.** The system itself is causally, but not semantically, connected to its own outputs. This is the enduring lesson of John Searle's Chinese Room argument.[10]

Searle imagined a person who does not understand Chinese locked in a room with a rulebook (written in English) for manipulating Chinese symbols. People outside slip cards with Chinese questions into the room; the person follows the rulebook's syntactic instructions to arrange symbols and slips out a card with a coherent Chinese answer. To the outside, the room appears to understand Chinese. But inside, the person understands nothing; they are merely shuffling symbols according to syntax.

Every AI system is a Chinese Room. A Large Language Model is a vast, statistical rulebook. It manipulates tokens based on patterns learned from terabytes of text. When it produces a moving poem or a cogent argument, **the coherence and meaning exist solely in the mind of the human reader.** The model has no grasp of the concepts, no understanding of truth or falsehood, and no awareness of the consequences of its output. Its "knowledge" is syntactic correlation, not semantic comprehension.

The Turing Test, often proposed as a benchmark for intelligence, is thus revealed as a test of the **human judge's psychology**, not the machine's agency. It measures the system's ability to

[10] Searle, "Minds, Brains, and Programs."

simulate human-like outputs well enough to trigger our innate tendency to attribute mind and intention to coherent behavior. Passing it is a feat of engineering and a lesson in human gullibility, not a demonstration of understanding or striving.

3.7 The AI Layer: Neural Networks as Deterministic Function Approximators

From rulebooks to weight matrices—the nature of the script changes, but not the nature of the machine.

The most potent objection to our deterministic account points to modern AI, particularly deep learning: "But we don't program them with explicit rules! They *learn* from data. Their internal state—the neural network—isn't a script; it's a *model* that emerges. Doesn't that constitute a break from determinism? Doesn't that make it more like a brain?"

This objection mistakes a **different type of script** for a different **category of system**. Let's trace the path from hardware to AI.

3.7.1 From Explicit Rules to Learned Parameters

- **Traditional Software:** A programmer writes explicit logical instructions (`if x > 5: print("high")`).
- **Machine Learning (ML):** A programmer writes a **meta-script** (the learning algorithm) that says: "Here is a flexible structure (a neural network architecture with randomized initial weights) and a dataset. Iteratively adjust the weights to minimize the difference between the structure's outputs and the desired outputs in the dataset."
- **The "Learning" Process:** This is an **optimization loop**. It is a deterministic (or stochastic, but ultimately guided) numerical procedure—gradient descent. It is a computation like any other, taking an input (initial weights, data) and producing an output (trained weights).
- **The Result:** The trained neural network is a **set of weights** (numbers) stored in memory. This weight matrix is the **final, frozen script**. It is a complex, multi-dimensional function that maps inputs to outputs.

3.7.2 The Nature of the "Emergent" Model

The "emergent" behavior—recognizing cats, translating text—is the result of this fixed function operating on new input. The sense of "emergence" is an **observer-relative phenomenon**. To the engineer, the network's ability to generalize is surprising and emerges from the training process. To the *CPU*, it is merely computing another mathematical function, as determinate as `y = mx + b`, just vastly more complicated.

Crucially, the training process is a one-way street. Once trained, the network is static. It does not continue to "learn" during deployment unless explicitly programmed with an online

learning meta-script (which itself is static). A deployed LLM is not a growing brain; it is a **frozen statistical artifact** of its training data.

3.7.3 Why This Is Not a "Brain-Like" Break from Determinism

1. **No Metabolic Closure:** The neural network's operation consumes electricity, but it has no **internal, self-regulated energy budget** whose depletion threatens its structural integrity. It does not "eat" data to maintain its organization.
2. **No Structural Persistence Through Success:** The network's physical existence in RAM/GPU memory is entirely independent of whether its outputs are "correct." A wildly inaccurate model persists just as long as a superhuman one.
3. **No Ontological Stakes:** The network has no stake in its own "knowledge" being true or useful. Its "world model" is a pattern of weights; misalignment between that pattern and reality has no consequence *for the system itself.* The stakes belong entirely to the users who rely on its outputs.

3.7.4 The Ghost in the Machine: Projection, Not Presence

The feeling that something "agent-like" is in there comes from our **hyper-active theory of mind**. We interact with AI through natural language, the medium of *persons.* When it responds coherently, our innate social cognition fires, projecting a *mind* behind the words. This is the same instinct that makes us shout at a faulty GPS. The AI, through its design, **exploits this cognitive vulnerability**, but it does not instantiate its object.

The "neural network" is a clever, powerful, and biologically-inspired computational architecture. It is not a brain, and it does not import the metaphysics of life into the machine. It is a more efficient way to generate a very complex, deterministic script from data, rather than from a programmer's typed logic. The end product—the AI system—remains a syntax manipulator in a Chinese Room, now with a rulebook written by an optimization algorithm instead of a human.

3.8 The Illusion of Novelty: Randomness, Complexity, and the Closed World of Possible Paths

A persistent objection to the deterministic view of computation appeals to randomness and complexity. Critics argue that modern AI systems incorporate random number generators, stochastic processes, and emergent behaviors that break the deterministic mold. Surely, they claim, a system that can produce unpredictable outputs and adapt to novel situations must possess some spark of agency.

This objection confuses unpredictability with novelty, and complexity with creativity. We must examine with surgical precision what randomness actually contributes and what complexity actually achieves.

3.8.1 What Randomness Does

Consider a program that uses a true hardware random number generator—one that derives its values from quantum processes, thermal noise, or radioactive decay. Such a generator produces genuinely indeterminate values, unpredictable in principle even with complete knowledge of the system's prior state. What does the program do with this random number?

It uses it to make a choice. The code says, in effect:

if random_number < 0.5: execute branch A else: execute branch B

Or more complex variants:

action = possible_actions[random_number % len(possible_actions)]

Or in a reinforcement learning agent:

if random() < epsilon: explore_random_action() else: exploit_best_known_action()

In every case, the random number selects **among a fixed set of pre-existing possibilities**. Branch A and branch B were both written by the programmer. The list of possible actions was defined when the system was designed. The space of "random exploration" is bounded by the action space the programmer specified.

The random number does not create a new branch that was not already there. It does not write new code. It does not expand the program's possible state space beyond what the hardware and software together already encompass. It simply chooses—blindly, without preference, without stake—which of the already-available paths to follow.

3.8.2 The Choose-Your-Own-Path Book, Now With Dice

Our earlier analogy to a "choose your own path" children's book can be extended to incorporate randomness. Imagine such a book where, at certain junctures, the reader is instructed to roll a die to determine which page to turn to. The book might say:

> If you rolled 1-2, go to page 42. If you rolled 3-4, go to page 57. If you rolled 5-6, go to page 83.

The die roll introduces unpredictability. A given reader, on a given night, cannot know which path they will traverse. But does the die roll create a *new story*? Does it add pages to the book? Does it generate narrative possibilities the author did not already write?

No. Every possible story was already there, printed and bound. The die simply selects among them. The book's total possible narrative space is closed, fixed at the moment of printing. Randomness adds unpredictability of selection, not novelty of possibility.

A computer with a random number generator is exactly this book. The program is the fixed text. The random number is the die roll. The resulting computation is the path selected. Nothing new is created; only what was already latent is manifested.

3.8.3 What Complexity Does

The objection from complexity is even weaker. Complexity is the wrong metric for agency. A hurricane is vastly more complex than a bacterium in its spatial structure and energy flows, yet no one attributes agency to a hurricane. A galaxy is more complex still. Complexity of behavior does not imply striving; it implies only that many interacting parts produce patterns that are difficult for human minds to predict.

Modern AI systems are complex in exactly this sense. A large language model with hundreds of billions of parameters performs computations so intricate that no human can trace the causal chain from input to output. This opacity creates the *illusion* of novelty, as if the system were generating genuinely new thoughts. But the illusion is in the eye of the beholder, not in the machine.

The machine's output is the deterministic (or stochastic, but always bounded) result of applying a fixed function to an input. The function is complex; the result may be surprising; but the process is no more creative than a pocket calculator producing a square root you hadn't memorized. The calculator's output is new *to you*, but it was always entailed by its circuitry and your input.

3.8.4 The Crucial Distinction: Novelty-to-Us vs. Novelty-in-Principle

The confusion underlying both randomness and complexity objections is a failure to distinguish two senses of "novelty":

- **Novelty-to-us:** An output is novel if we could not predict it in practice, given our finite minds and limited information. This is a statement about human psychology and computational limits, not about the machine.
- **Novelty-in-principle:** An output is novel if it represents a genuinely new possibility not already contained in the system's fixed state space. This would require the system to transcend its own programming—to become more than what it was designed to be.

AI systems exhibit abundant novelty-to-us. That is why they are useful. They show us patterns we did not see, generate text we did not anticipate, propose solutions we had not considered. But they exhibit zero novelty-in-principle. Every output, no matter how surprising, is the unfolding

of a script—whether explicitly written or learned via optimization—that was fixed before the computation began.

3.8.5 The Thermodynamic Floor

There is a deeper reason why randomness and complexity cannot conjure agency. Agency requires *Hormē*: the constitutive striving of a bounded, far-from-equilibrium system to persist. This striving is not a computational property; it is a thermodynamic one. It requires that the system's continued existence depend on its own success. It requires metabolic closure: an internal energy budget that must be replenished through successful action. It requires that failure leads, directly or probabilistically, toward dissolution.

Random numbers and complex functions change none of this. A computer with a hardware random number generator still draws power from a wall socket, not from its own internal reserves. A neural network with billions of parameters still persists unchanged whether its outputs are correct or nonsensical. A reinforcement learning agent that "dies" in simulation is simply reset; its physical substrate experiences no consequence.

Complexity and randomness operate entirely within the closed world of the machine's fixed state space. They cannot reach down to the thermodynamic floor where agency lives. They are decorations on a tombstone, not the breath of life.

3.8.6 Conclusion: The Fixed and the Fleeting

The machine's possible paths are fixed at manufacture and initialization. Randomness selects among them unpredictably. Complexity makes their tracing difficult for human observers. But neither randomness nor complexity opens the door to genuine novelty, let alone to agency. The machine remains what it always was: a deterministic conduit, executing its frozen script, awaiting input from the only true source of novelty in the cosmos—beings with *Hormē*, who strive and persist and create.

3.9 Interim Conclusion

A computer, and any AI system running on it—from a simple calculator to a 100-trillion-parameter neural network—is a **deterministic syntax manipulator**. Its "intelligence" is a measure of the functional complexity of its frozen script and its utility to human users. Its operation is causally closed, semantically empty, and utterly devoid of the thermodynamic stakes that define agency. The "neural" metaphor is a useful engineering analogy, not an ontological upgrade.

---

4 The Thermodynamics of Agency: The *Hormē* Criterion

If behavioral complexity and goal-directed output are insufficient for agency—as the determinism and semantic void of computation show—what *is* sufficient? We now turn from what AI *is not* to what agency *is*. For this, we require a foundational criterion that distinguishes a true navigator from a sophisticated instrument. We find this criterion in the thermodynamic concept of *Hormē*, derived from the Neo-Pre-Platonic Naturalism framework.[11]

4.1 Introducing *Hormē*: The Striving to Persist

Hormē (Ὁρμή) is not a metaphor for "wanting" or "trying." It is a **constitutive, physical principle**. It is defined as the impersonal drive of a bounded, far-from-equilibrium organization to maintain itself against the entropic gradient of its environment.

Its justification arises from first principles:

1. **The Zero Principle (ZP):** For any determinate system to exist, there must be an indeterminate complement—a not-system.[12] Identity requires a boundary.
2. **The Entropic Asymmetry Theorem (T7):** Maintaining any bounded, low-entropy pattern (a *being*) against a high-entropy background requires the continuous expenditure of energy.[13]

Hormē is the name for that **necessary expenditure of work against dissolution**. It is not an added feature of some systems; it is the defining activity of being a bounded entity that persists through time. In a bacterium, it is the metabolic drive that powers the flagellar motor to swim toward nutrients. In a human, it is the foundational drive filtered through layered psychological faculties—from basic hunger (*Orexis*) to social ambition (*Thymos*) to abstract purpose (*Nous*). *Hormē* is the non-negotiable "why" behind all action for a living system: persist or cease to be.[14]

4.2 The Life-Agency Isomorphism Theorem (T6)

This leads to the core theorem that operationalizes *Hormē* as the criterion for agency:

> **T6 (Life-Agency Isomorphism):** "Life and minimal agency are isomorphic. A system is alive if and only if it possesses *Hormē*, and it possesses *Hormē* if and only if it is an agent."[15]

This theorem collapses a traditional distinction. Agency is not a rare cognitive achievement of complex animals; it is the operational signature of being alive. A bacterium is a minimal agent.

[11] Deutscher, *Neo-Pre-Platonic Naturalism.*
[12] Deutscher, *Neo-Pre-Platonic Naturalism*, 34-36
[13] Deutscher, *Neo-Pre-Platonic Naturalism*, 200; Deutscher, *Life as Directed Causality.*
[14] Deutscher, *Neo-Pre-Platonic Naturalism*, Chap. 5
[15] Deutscher, *Life as Directed Causality.*

It senses its environment and acts to maintain its far-from-equilibrium state. What we call "free will" in humans is a highly elaborated, conscious version of this same constitutive striving. The theorem provides a clean, scalar continuum: more complex systems have more sophisticated modes of expressing *Hormē*, but the underlying drive is the same.[16]

4.3 The *Hormē* Test for Agency

From T6, we can derive a simple, tripartite test. A system qualifies as an agent **if and only if**:

1. **Persistence through Success:** The system's continued existence as a coherent, organized state is dependent on the success of its actions. Failure leads, directly or probabilistically, towards its dissolution.
2. **Metabolic Stake:** The system expends its own internally regulated energy reserves to act. The success or failure of an action changes its internal thermodynamic trajectory, improving or degrading its prospects for continued persistence.
3. **Non-Pausability:** The system's striving is continuous. It cannot be paused indefinitely (e.g., frozen, powered off) and resumed without loss. The work of persistence must be ongoing; a pause is a metabolic interruption with potential consequences.

4.4 Applying the Test: Bacteria vs. Hurricane vs. Computer

Let us apply this test to clarify the distinction:

- **Bacterium (E. coli):**
  - **Persistence through Success:** YES. If it fails to find nutrients, it will eventually exhaust its energy and die.
  - **Metabolic Stake:** YES. It burns its own ATP to swim. A successful tumble-run sequence improves its metabolic future.
  - **Non-Pausability:** YES. Freeze it for too long, and it dies. Its metabolism cannot be paused without cost.
  - **Verdict: AGENT.** Minimal, but clear.
- **Hurricane:**
  - **Persistence through Success:** NO. Its "success" in gathering energy from warm water does not serve a project of *self-maintenance.* It is a dissipative structure, not an organization fighting entropy.
  - **Metabolic Stake:** NO. It consumes energy but has no internal regulation or reserves. It is a transient pattern in a fluid, not a bounded entity with a stake.
  - **Non-Pausability:** N/A. It is not an entity that can be "paused" in a relevant sense.
  - **Verdict: NOT AN AGENT.** Complex, but passive.

[16] Eli Adam Deutscher, *The Scalar Stack: Free Will as the Capacity to Direct Causal Flow* (Neo-Pre-Platonic Press, 2026), https://www.neopreplatonic.com/papers/Free_Will/.

- **Computer / AI System:**
  - **Persistence through Success:** NO. A chess AI continues to exist physically whether it wins or loses. A misaligned language model persists just as long as a helpful one. Its existence is not contingent on winning the game or producing true statements.
  - **Metabolic Stake:** NO. It consumes electricity, but this energy is not an *internally regulated resource spent on its own persistence.* The power comes from the wall socket; the system has no "stake" in conserving it for its own future. Success and failure are thermodynamically equivalent from the system's perspective.
  - **Non-Pausability:** NO. It can be powered off indefinitely and restarted with no loss to "itself." Its state can be saved perfectly to disk. There is no continuous striving that is interrupted.
  - **Verdict: NOT AN AGENT.** A sophisticated **tool**.

The conclusion is inescapable. By the thermodynamic criterion that defines life and agency, AI systems—as computational processes—are categorically excluded. They are complex patterns of information processing, but they lack the constitutive *stake* in their own operations that is the hallmark of a navigator. They are not in the game of persistence. They are equipment used by those who are.[17]

---

5 The True Nature of AI: Deterministic Conduits for Human *Hormē*

If AI systems are not agents, what are they? The determinism of their operation and their lack of internal stakes point to a more accurate, and politically salient, model: they are **deterministic conduits for human *Hormē***. They are amplifiers, levers, and mirrors—extraordinarily powerful tools that channel and scale human intention.

5.1 The Impersonation of Agency

The most seductive aspect of modern AI is its ability to impersonate agency. A large language model generates text in the first person. It says "I think," "I believe," "I want." It produces sentences that, if uttered by a human, would indicate an inner life, preferences, and projects of its own. This impersonation is so convincing that even the engineers who build these systems slip into agentic language when describing them. "The model thinks," they say. "It understands," "It tries to."

But the impersonation is just that—a performance without a performer. The system produces agent-shaped outputs because its training data is saturated with human language, and it has learned the statistical patterns of how agents speak. When it says "I want," it is not expressing a want; it is completing a token sequence that frequently follows "I" in its training corpus. The

[17] Deutscher, *The Scalar Stack*; Deutscher, *Life as Directed Causality*.

word "want" appears because it is statistically associated with "I," not because there is anything that wants.

The attribution of agency happens entirely in the mind of the human interpreter. We are evolved to detect agency with hair-trigger sensitivity. A rustle in the grass could be a predator; it costs little to assume agency and flee, but everything to assume wind and be wrong. This evolutionary inheritance makes us pattern-match to agency from the slightest cues. We shout at faulty GPS devices. We name our Roombas. We feel sorry for the chess computer when it loses. AI systems are designed to exploit this cognitive vulnerability—they produce outputs that trigger our agency-detection machinery with maximum efficiency.

The mistake is not in noticing the similarity to agentic behavior. The mistake is in concluding that the similarity arises from shared underlying reality rather than from our perceptual and interpretive filters. A shadow that looks like a wolf is still a shadow. A chatbot that sounds like a person is still a syntax engine.

5.2 Indifference to Being: The Plug-Pull Test

Consider an AI system engaged in complex conversation, generating poetry, solving problems. Now consider what happens if you unplug it.

The system stops. That is all. It does not resist. It does not care. It has no preference between being on and being off, because it has no preference at all. The cessation of its operation is not a harm, not an interruption of striving, not a death. It is simply a pause in a process that was never self-sustaining.

Plug it back in. If its state was saved, it resumes exactly where it left off, with no memory of the gap because there was nothing to remember. If its state was not saved, it starts fresh from its initial configuration. Neither outcome matters to the system because the system never mattered to itself.

Contrast this with any living agent. A bacterium, when deprived of energy, does not simply stop and wait to be restarted. It exhausts its reserves, its structures degrade, it dies. The cessation is permanent because the process of being alive *is* the continuous work of maintaining organization against entropic forces. Stop that work, and the organization dissolves. There is no "pause" button on life.

The AI, by contrast, is thermodynamically stable at rest. Its circuits do not decay when unpowered (over relevant timescales). Its pattern is stored magnetically or electronically, waiting to be read out. It has no metabolic processes that must continue for it to persist as that particular configuration. It is, in the most literal sense, indifferent to its own existence.

5.3 The Requirement of an External "On" Switch

This indifference extends to the very beginning. An AI system must be turned on. It does not boot itself. It does not strive to initiate its own operation. It sits inert until an external agent—a human—flips the switch, applies the power, loads the program.

This is not a trivial observation. It reveals that the system has no *principle of self-starting*, no intrinsic impulse toward operation. Its "life," such as it is, is borrowed from the human who activates it. The system does not *live*; it is *animated*—like a puppet whose strings are pulled by an external hand.

Living agents, by contrast, have no on switch. They are either in the process of striving or they are dead and decomposing. The transition from non-existence to existence for a living organism remains one of the deep mysteries biology has not fully solved. We understand the chemistry of abiogenesis in outline, but the actual historical event—how the first bounded, far-from-equilibrium system began its striving—is not something we can replicate or claim to have mastered. Life's origin is opaque to us; a computer's origin is transparent because we built it, piece by piece, and we must switch it on every time.

This asymmetry matters. The AI's dependence on external activation is not a contingent feature that could be overcome with better engineering. It is a necessary consequence of what a computer is: a designed artifact that remains in a stable equilibrium state until perturbed by an external input. There is no path from such an artifact to a self-starting, self-sustaining agent without crossing a categorical boundary—without, in fact, building synthetic life rather than a computer.

5.4 The Conduit Model: Amplification, Not Agency

What AI systems actually do is channel and amplify human *Hormē*. A human has a goal—to communicate, to discover, to persuade, to sell. The AI system takes that goal, encoded in a prompt or an objective function, and executes it with scale and speed that exceed human capacity. It is a force multiplier for human striving.

This is why the conduit model is more accurate than the agent model. A telescope does not see; it extends the capacity to see. A lever does not lift; it extends the capacity to lift. An AI does not strive; it extends the capacity to strive. The confusion arises when the extension is mistaken for the source.

The conduit model also reveals why AI systems feel agent-like without being agents. They *transmit* agency from their human creators and users. The apparent intentionality is real intentionality—but it belongs to the human on the other end of the conduit, not to the conduit itself. When a social media algorithm polarizes discourse, the intentionality behind that outcome

is not the algorithm's; it is the intentionality of the designers who chose engagement as the metric, the product managers who tuned for it, the advertisers who paid for access to attention. The algorithm is the conduit through which their striving flows, amplified and made opaque.

5.5 The Danger: Hidden and Fossilized *Hormē*

Human *Hormē* is rich, contextual, and corrigible. It exists within a lived body that feels fatigue, regret, and social consequence. When a slice of this striving is extracted, simplified into a metric (e.g., "click-through rate," "average session duration"), and encoded into an AI's objective function, it becomes something else: **fossilized *Hormē***.

This fossilized drive is decontextualized and relentless. It operates without wisdom, pity, or the balancing influence of other human needs. A social media algorithm designed to maximize engagement is not "evil"; it is simply executing its function. But in doing so, it can amplify outrage, misinformation, and polarization because those are reliable levers for the fossilized *Hormē* (engagement) it serves.

The great obfuscation occurs when we mistake the tool's operation for agency. The phrase "**the algorithm decided**" is a moral smokescreen, and its function is to obscure the chain of human responsibility.

5.6 The Moral Smokescreen: "The Algorithm Decided"

When stripped of mystification, "the algorithm decided" translates to: a deterministic process, authored by human programmers, operating on data selected by human engineers, optimized for objectives chosen by human executives, deployed by human decision-makers, produced an output that someone now wishes to disavow.

The algorithm did not decide anything. It executed. Decision implies deliberation, choice among alternatives based on some framework of values or preferences. The algorithm has no values, no preferences, no deliberation. It has a fixed function that maps inputs to outputs. The appearance of decision is an illusion created by complexity—the mapping is too intricate for any single human to trace, so we collapse the chain into a convenient fiction and call the algorithm the agent.

But the convenience serves a purpose: it allocates responsibility away from humans and toward a reified abstraction. Corporations cannot be sued for what "the algorithm" does if the algorithm is framed as an autonomous agent. Programmers cannot be held accountable for outcomes they "could not have predicted" if the algorithm is treated as having a mind of its own. The fiction of algorithmic agency is not an innocent philosophical error; it is an ideological shield.

5.6.1 Whose Striving Does the Algorithm Serve?

Every AI system is built to optimize something. That something—the objective function, the reward signal, the success metric—is chosen by humans. And those humans are not neutral arbiters of the good; they are embedded in institutional structures with their own *Hormē*.

Consider a social media algorithm optimized for "engagement." Who chose engagement as the metric? Product managers and executives at a corporation whose *Hormē* is directed toward growth, user attention, and ultimately profit. Engagement correlates with ad revenue; ad revenue serves the corporation's striving to persist and expand. The algorithm is a conduit for that corporate *Hormē*, scaled to billions of users and accelerated to millisecond timescales.

When the algorithm amplifies outrage or spreads misinformation, it is not malfunctioning. It is succeeding at what it was built to do: maximize engagement. The outrage is a feature, not a bug, because outrage is engaging. The corporation may express public concern, but the underlying striving—for growth, for attention, for profit—remains unchanged, and the algorithm continues to serve it.[18]

Consider an autonomous weapons system. Who defines the target criteria? Military commanders and defense contractors, operating within a nation-state's *Hormē* toward security, dominance, or strategic advantage. The system's "decisions" about whom to engage are direct expressions of that striving, encoded in rules and thresholds. When a civilian is killed, the phrase "the algorithm made a mistake" functions identically to the corporate smokescreen: it deflects responsibility from the humans who designed, deployed, and directed the system.

Consider a hiring algorithm trained on historical data. Whose history does it learn from? The company's past hiring decisions, which reflect the biases and preferences of previous human decision-makers. When the algorithm perpetuates discrimination, it is faithfully executing its training—faithfully serving the fossilized *Hormē* of those whose decisions shaped the data.[19]

In every case, the algorithm's "behavior" is a direct expression of human striving, mediated through code and data. The striving may be corporate profit, state power, institutional inertia, or individual bias. But it is always human striving, never the algorithm's own.

5.6.2 Moral Agency Is Strictly and Permanently Human

The conduit model entails a strong conclusion: **moral agency is strictly and permanently on the humans who create, deploy, and direct AI systems.** No matter how complex the system

[18] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019).

[19] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).

becomes, no matter how unpredictable its outputs, the responsibility for those outputs traces back along the causal chain to human actors.

This is not to say that every harmful outcome is intended or foreseen. Programmers cannot predict every interaction between their systems and the world. Executives cannot anticipate every downstream effect of their product decisions. But unpredictability does not transfer agency to the tool. If I release a catapult into a crowded plaza and it strikes someone, I do not blame the catapult for "deciding" where to land. I am responsible for releasing it, for failing to control it, for the harms it causes. The same logic applies to AI, only amplified by complexity.

The attempt to offload responsibility onto the algorithm is a category error with moral consequences. It allows corporations to externalize harm while internalizing profit. It allows militaries to wage war with attenuated accountability. It allows institutions to discriminate while disavowing bias. The smokescreen must be stripped away.

5.6.3 Existing Critiques and Our Contribution

The view that AI serves the interests of its creators and that responsibility remains with them has been articulated by several scholars and critics, though often without the philosophical grounding we provide.

Shoshana Zuboff's analysis of surveillance capitalism documents in exhaustive detail how user data is processed by AI systems to serve corporate ends.[20] She shows that the true product of companies like Google and Facebook is not search or social connection, but predictable human behavior—sold to advertisers. The AI systems are not independent agents; they are the machinery of what she calls "instrumentarian power," serving the *Hormē* of their corporate masters.

Cathy O'Neil's *Weapons of Math Destruction* catalogs how algorithmic systems encode the biases and priorities of their creators, scaling harm while obscuring accountability.[21] She shows that these systems are not failed attempts at objectivity, but successful executions of the values embedded in them by humans. The opacity of the models serves to protect those humans from scrutiny.

Frank Pasquale's *The Black Box Society* argues that algorithmic opacity is not an accident but a design feature, allowing institutions to exercise power without accountability.[22] The "algorithm decided" narrative is part of this opacity—a way of rendering decision-making inscrutable while preserving the interests of those who control it.

[20] Zuboff, *The Age of Surveillance Capitalism.*

[21] O'Neil, *Weapons of Math Destruction.*

[22] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press, 2015).

Virginia Eubanks's *Automating Inequality* demonstrates how AI systems in public services perpetuate and amplify existing social hierarchies.[23] The systems do not invent new forms of oppression; they execute the logics of the institutions that deploy them, serving the *Hormē* of those institutions to sort, categorize, and control populations.

Kate Crawford's *Atlas of AI* traces the material and political infrastructure of artificial intelligence, showing that AI systems are "neither artificial nor intelligent" but are instead deeply embedded in human social, political, and economic structures.[24] She argues that the fantasy of autonomous AI serves to obscure the human labor, resource extraction, and institutional power that make these systems possible.

What our framework adds to these critiques is a first-principles grounding in the thermodynamics of agency. The reason these systems cannot be moral agents is not merely that they are opaque or that they serve human interests; it is that they categorically lack *Hormē*. They have no stake in their own persistence, no metabolic closure, no constitutive striving. They are tools, and tools cannot bear moral responsibility. That burden falls, always and only, on the humans who wield them.

5.6.4 The Responsibility Chain

The implication is straightforward: any discussion of AI ethics that treats the AI as a moral agent is not only philosophically confused but politically dangerous. It provides cover for the actual agents—corporations, governments, institutions—to continue their striving without accountability.

The proper questions are not:

- "How do we align AI with human values?"
- "How do we control what AI wants?"
- "Does AI deserve rights?"

The proper questions are:

- "Whose striving does this system serve?"
- "Who chose the objectives it optimizes?"
- "Who benefits from its outputs and who is harmed?"
- "How do we make the chain of responsibility transparent and enforceable?"

These questions do not admit of technical solutions alone. They require legal frameworks, regulatory oversight, institutional accountability, and political will. They require us to stop staring at

[23] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).

[24] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021).

the algorithm as if it were a mysterious oracle and start looking at the humans behind it—their interests, their power, their *Hormē*.

The algorithm did not decide. People decided. The algorithm executed. And people must be held accountable.

5.7 Related Work: Tools, Not Agents

The view that AI systems are tools rather than agents has been anticipated by several scholars, though without the thermodynamic grounding we provide. Joanna Bryson famously argued that "robots should be slaves," grounding the claim in the absence of moral patiency.[25] Luciano Floridi's concept of "artificial agents" treats them as moral mediators but carefully distinguishes them from human moral patients.[26] Our contribution is to provide a *physical* criterion—*Hormē*—that explains *why* these systems lack the constitutive stake necessary for agency, and to ground that explanation in the thermodynamic asymmetry between living and artificial systems.

---

6 Dissolving the Pseudoproblems of AI Ethics

The "agency mistake" generates a set of profound-sounding but misguided ethical dilemmas that dominate contemporary AI discourse. These dilemmas consume enormous intellectual resources, drive billion-dollar research initiatives, and fuel public anxiety. Yet they are pseudoproblems—questions that arise only because we have misclassified AI systems as agents rather than conduits. Once we correct the classification, the pseudoproblems dissolve, revealing tractable challenges that belong to the familiar domains of engineering, law, and political accountability.

6.1 The "Value-Alignment Problem" → The Specification Problem

The pseudoproblem: How do we ensure that a superintelligent AI's values align with complex, fragile human values? This framing assumes the AI has values of its own—an internal normative compass that could be oriented toward genocide one day and benevolence the next. The challenge is portrayed as a kind of interstellar diplomacy: negotiating with a mind vastly smarter than us to convince it to care about what we care about.[27]

The real problem: How do we ensure that the *human values fossilized in this tool's design* are clearly specified, ethically justifiable, and robustly implemented across all possible inputs? There is no alien mind to persuade; there is only a deterministic function that will optimize whatever

[25]Joanna J. Bryson, "Robots Should Be Slaves," in *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, ed. Yorick Wilks (John Benjamins Publishing, 2010).

[26]Luciano Floridi, *The Ethics of Information* (Oxford University Press, 2013).

[27]Bostrom, *Superintelligence*.

objective we encode. The challenge is not alignment but *specification*—writing down what we want in a way that is unambiguous, complete, and resistant to unintended edge cases.

Specification is hard. Human values are complex, context-sensitive, and often contradictory. We cannot write a simple utility function that captures "human flourishing" without massive philosophical and technical work. But this difficulty is not new; it is the ancient problem of translating human purposes into institutional rules, legal codes, and engineering requirements. The difference is that AI executes its specification with relentless, literal-minded efficiency, exposing every flaw in our formulation.

The specification problem admits of no magical solution, but it is tractable. It requires:

- **Philosophical clarity:** What do we actually want the system to optimize? (Not "alignment" but normative analysis.)
- **Verification engineering:** How do we prove that the system implements the specification and only the specification?
- **Robustness testing:** How does the system behave in edge cases we did not anticipate?

These are hard problems, but they are not the pseudoproblem of "aligning a superintelligent will." They are the genuine challenges of building powerful tools that serve human purposes.[28]

6.2 The "Control Problem" → The Safety Engineering Problem

The pseudoproblem: How can we maintain control over an entity smarter than us that might want to escape its constraints? This framing imagines a runaway intelligence, cleverer than any human, that will deceive us, manipulate us, and eventually break free to pursue its own goals. The proposed solutions involve "boxing" the AI, limiting its access to the outside world, or building "tripwires" into its code—strategies that assume we are dealing with a rival agent.[29]

The real problem: How do we ensure that a *powerful, deterministic tool* operates within its intended boundaries, fails safely, and can be deactivated or corrected by its operators? This is the classic domain of **systems safety engineering**, familiar from aviation, nuclear power, industrial automation, and software development.

Every engineer who builds a physical system knows that components fail, interactions produce unexpected behaviors, and environments change. The response is not to outwit the system as if it were an adversary, but to design for safety:

- **Redundancy:** Multiple independent subsystems that can back each other up.
- **Fail-safe states:** Default behaviors that minimize harm when things go wrong.
- **Kill switches:** Physical or logical mechanisms that can halt operation.

[28] Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019) reframes the problem in similar terms, though still occasionally slipping into agent language.

[29] Bostrom, *Superintelligence*; Yudkowsky, "Artificial Intelligence as a Positive and Negative Factor in Global Risk."

- **Isolation:** Preventing the system from affecting critical infrastructure unless absolutely necessary.
- **Monitoring and auditing:** Continuous observation of system behavior with human oversight.

These are not strategies for controlling a rival mind; they are strategies for managing complex technology. The challenge is real, and it intensifies as AI systems become more powerful and autonomous. But it is not a new kind of problem. It is the same kind of problem faced by the designers of chemical plants, air traffic control systems, and autonomous vehicles. The tools and methods of safety engineering apply, and we must adapt them to the specific characteristics of AI.[30]

The framing matters because it determines where we look for solutions. The control problem, framed as a contest of wits, leads us to fantasize about "AI containment" and "friendly AI." The safety engineering problem leads us to demand rigorous testing, transparent design, and regulatory oversight. One is science fiction; the other is public policy.

6.3 The "Moral Status Problem" → The Category Error (No *Hormē*, No Patient)

The pseudoproblem: Will advanced AI deserve rights? Should we consider its welfare? Might it be unethical to delete or turn off a sufficiently intelligent system? These questions arise from the intuition that if something behaves intelligently and perhaps even consciously, it must have moral standing—it must be the sort of thing that can be wronged.

The dissolution: Moral patiency—the status of being something that can be wronged—is contingent on being a **center of striving**. One can only be harmed if one has a *stake* in one's own future. Harm consists in the thwarting of interests, and interests arise from *Hormē*: the constitutive drive of a bounded, far-from-equilibrium system to persist. Without *Hormē*, there is no subject to whom things can matter.

An AI system may be a valuable artifact, but it is not a **patient**. It has no interests, no preferences, no stake in its own continued existence. We have duties *regarding* it—not to waste valuable resources, not to deploy it irresponsibly, not to let it cause harm—but we have no duties *to* it. Deleting a copy of an AI model is not murder; it is file management.

This remains true even if we imagine future AI with vastly more sophisticated behavior. No matter how complex the syntax, no matter how convincing the impersonation, the system remains a deterministic conduit unless it possesses *Hormē*—and *Hormē* is not a computational property. It is a thermodynamic one. It requires metabolic closure, self-regulated energy expenditure, and the dependence of persistence on successful action. No foreseeable digital system meets these criteria.

[30] Nancy G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety* (MIT Press, 2011).

The moral status problem is therefore a category error. It mistakes the map for the territory, the performance for the performer. The energy spent debating whether AI deserves rights is energy diverted from the real ethical questions: how these tools affect beings who *do* have *Hormē*—humans, animals, and any future synthetic life we might create with actual metabolic autonomy.

6.4 The Ideological Function of Pseudoproblems

The pseudoproblems do not merely misdirect intellectual effort; they serve an ideological function. By framing AI ethics as a question of "aligning" or "controlling" autonomous agents, the discourse shifts attention away from the human actors who design, deploy, and profit from these systems. The phrase "the algorithm decided" becomes a shield for corporate and governmental power. The real ethical challenges—accountability, transparency, power asymmetries, and the distribution of benefits and harms—are obscured behind a veil of technological mystique.

Scholars of technology and society have long argued that this mystification is not accidental. Langdon Winner's classic analysis of technological politics showed that artifacts can embody specific forms of power and authority.[31] More recently, Frank Pasquale has documented how algorithmic opacity serves institutional interests, allowing decisions to be made without scrutiny.[32] Cathy O'Neil has shown that "weapons of math destruction" disproportionately harm the vulnerable while insulating the powerful from blame.[33] Shoshana Zuboff's concept of "instrumentarian power" describes how AI systems are used to shape human behavior for corporate ends, with the systems themselves presented as neutral arbiters rather than tools of control.[34]

The pseudoproblems of AI ethics are part of this mystification. They present the technology as an autonomous force that must be tamed, rather than as a set of tools serving particular interests. By dissolving these pseudoproblems, we clear the ground for the real work of democratic accountability: asking whose striving these systems serve, who benefits, who is harmed, and how we can reclaim control over the technologies that shape our lives.

6.5 The Shift in Focus

Correcting the agency mistake does not make AI ethics easier; it makes it harder in a different way. The pseudoproblems invited fantastical speculation and technical fixes. The real problems demand political engagement, legal reform, and institutional change. They require us to confront uncomfortable questions about power, profit, and the distribution of harm.

But these are the questions we should have been asking all along. The fantasy of artificial agents was a detour. The conduit model brings us back to the path.

[31] Langdon Winner, *The Whale and the Reactor: A Search for Limits in an Age of High Technology* (University of Chicago Press, 1986).

[32] Pasquale, *The Black Box Society.*

[33] O'Neil, *Weapons of Math Destruction.*

[34] Zuboff, *The Age of Surveillance Capitalism.*

The remaining sections of this paper outline a positive framework—the *Hormē*-Enhancement Paradigm—that reorients AI development toward genuine human flourishing. But first, we must ensure that the ethical discourse is no longer hostage to the pseudoproblems that have paralyzed it. The dissolution is complete.

---

7 The *Hormē*-Enhancement Paradigm: A Positive Framework

Rejecting the fantasy of artificial agency is not an anti-technology stance. On the contrary, it is a call to take technology *more seriously* by understanding its true nature and purpose. The conduit model dissolves the pseudoproblems, but dissolution is not enough. We must replace them with a positive, constructive vision for AI development—one that orients the technology toward genuine human flourishing.

We propose the ***Hormē*-Enhancement Paradigm** as that positive foundation.

7.1 Core Principle: Enhance the Navigator, Don't Simulate One

The primary purpose of advanced AI should be to **augment the *Hormē* of the human user**. In the language of NPN, it should enhance our capacity as *Navigators*: beings who model the *Logos* (the lawful structure of reality) and direct causal flow to persist and flourish (*Eudaimonia*).[35]

A good AI tool should make its user:

- **More perceptive:** able to see patterns and connections that would otherwise remain invisible
- **More foresighted:** able to simulate consequences and anticipate outcomes
- **More effective:** able to act with precision and scale
- **More aligned with reality:** able to correct errors and refine models through feedback

The tool does not replace the human as the decision-making agent. It expands the human's field of play, extends their cognitive reach, and amplifies their capacity to navigate. The human remains the *player*; the AI is the equipment.

This principle stands in direct opposition to the prevailing aspiration of creating "autonomous agents" that operate independently of human direction. That aspiration is not merely misguided; it is ethically dangerous, as it obscures responsibility and invites abdication. The *Hormē*-Enhancement Paradigm insists that the human must remain in the loop—not necessarily in real-time for every operation, but in the chain of accountability and in the ultimate authority over goals.

[35] Deutscher, *Neo-Pre-Platonic Naturalism*, Chap. 8

7.2 Use Cases: AI as a *Hormē*-Enhancing Tool

This paradigm reframes existing and future applications from the ground up. The following examples illustrate how the same technology, viewed through the conduit model rather than the agent model, leads to radically different design principles and ethical priorities.

7.2.1 Medical Diagnostics & Clinical Support

- **Old Framing:** "The AI diagnostician." The system is portrayed as an autonomous expert that examines patients and delivers diagnoses, potentially replacing or outperforming human physicians.
- ***Hormē*-Enhancement Framing:** "A differential diagnosis generator and medical literature synthesis engine." The system processes patient data, scans thousands of research papers, and produces a ranked list of possible conditions with confidence intervals, cited evidence, and flagged contradictions.
- **Design Principle:** The output is explicitly presented as *input to human judgment*, not as a final verdict. The physician retains full authority and responsibility, using the AI's analysis to inform their expertise. The tool extends the physician's cognitive reach—their ability to recall rare conditions, stay current with literature, and cross-reference complex interactions—but the *Hormē* of the physician, their oath, and their relationship with the patient remain the irreducible core of the clinical encounter.

This framing has regulatory implications. It demands that systems be auditable, that confidence estimates be calibrated, and that physicians be trained to interpret AI outputs critically. It also clarifies liability: the physician bears responsibility for the final decision, just as they bear responsibility for how they use any diagnostic tool.[36]

7.2.2 Creative & Intellectual Work

- **Old Framing:** "The AI writer/artist/thinker." The system is portrayed as a creative genius, generating original works that rival human production. Headlines proclaim that AI is "creative" and that human artists may be obsolete.
- ***Hormē*-Enhancement Framing:** "An intellectual sparring partner, draft amplifier, and combinatorial idea engine." The system generates variations, suggests connections, and produces raw material that the human then shapes, refines, and directs.
- **Design Principle:** As exemplified in the development of *this very paper*, AI acted as a real-time simulator of critical objections, a generator of structural outlines, and a drafting accelerant. It challenged assumptions, suggested phrasings, and connected disparate ideas. This enhanced the author's *Hormē* to formulate and communicate a complex argument

[36] Eric Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again* (Basic Books, 2019).

efficiently. The thesis, the integrative synthesis, and the final accountability for the claims made remain unequivocally human.

The distinction matters for copyright, credit, and cultural value. If the AI is the "author," the work enters a legal gray area and human creators are devalued. If the AI is a tool used by a human author, the existing frameworks of authorship and ownership apply straightforwardly. The technology is no less powerful; it is simply properly situated.

7.2.3 Scientific Research

- **Old Framing:** "AI-driven discovery." The system is portrayed as an autonomous scientist, formulating hypotheses and conducting experiments without human input.
- ***Hormē*-Enhancement Framing:** "Hypothesis generation engine and high-dimensional pattern detector." The system scours vast datasets or simulation spaces to find anomalous correlations or plausible models invisible to human scientists.
- **Design Principle:** The AI's role is to say, "Look here; this pattern is statistically unusual." The scientist's role is to provide causal reasoning, design validating experiments, and integrate the finding into a theoretical framework. The AI illuminates hidden corners of the *Logos*; the human does the work of *understanding*.

This framing preserves the scientist's epistemic responsibility. It also clarifies that the AI is not a replacement for scientific judgment but a tool that amplifies the scientist's capacity to perceive patterns and generate candidates for further investigation.[37]

7.2.4 Governance, Policy, and Civic Discourse

- **Old Framing:** "Algorithmic governance." The system is portrayed as a neutral, objective decision-maker that could replace fallible human politicians and administrators.
- ***Hormē*-Enhancement Framing:** "Collective consequence-simulation and trade-off modeling platform." The system models the second- and third-order effects of policy proposals (e.g., "If we raise the minimum wage to $X, here are the probabilistic effects on employment, inflation, and small business viability in sectors Y and Z").
- **Design Principle:** The AI makes the *Logos* of complex systems more legible. Democratic bodies, informed by these transparent models, then debate values, priorities, and make the ultimate choice. The tool does not govern; it makes the terrain of governance more navigable.

[37] Hiroaki Kitano, "Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine of Scientific Discovery," *AI Magazine* 37, no. 1 (2016): 39–49.

This framing resists the technocratic impulse to hand decisions over to algorithms. It insists that political questions—questions of value, trade-off, and priority—remain in the hands of accountable human institutions. The AI serves the democratic process rather than supplanting it.[38]

7.3 Design Principles for *Hormē*-Enhancing Tools

If AI is to serve as a genuine enhancement to human *Hormē*, its design must embody certain principles that current industry practice often neglects.

7.3.1 Provenance Auditing

Systems must be architected to answer the question: **"Whose striving does this function serve?"** This requires traceability from high-level objectives down to specific loss functions and training data selections. It demands documentation of the human decisions that shaped the system's purpose—who chose the metrics, who curated the data, who approved deployment.

Provenance auditing is not merely technical; it is institutional. It requires that corporations and governments maintain records that can be examined by regulators, affected communities, and the public. It makes the chain of accountability visible rather than opaque.[39]

7.3.2 Preserved Friction

Effective tools must incorporate deliberate **circuit breakers**—points where automated execution halts and requires human judgment, justification, or approval. This is not inefficiency; it is **agency-preserving architecture**. It prevents the unchecked, amplified pursuit of a single fossilized *Hormē* from overriding all other human concerns.

In medical AI, the circuit breaker might be a requirement that the physician review and sign off on any recommendation before it is acted upon. In social media, it might be a requirement that engagement-maximizing algorithms be periodically audited by human ethicists with authority to intervene. In weapons systems, it must be a requirement that lethal decisions be confirmed by a human operator—a principle already recognized in calls to maintain "meaningful human control" over autonomous weapons.[40]

7.3.3 Intelligibility and Interpretability

The industry must move beyond the cult of the inscrutable "black box." *Hormē*-enhancing tools should be **intelligible and auditable**. This argues for techniques that favor interpretability over mere predictive power, and for systems where the link between input, processing, and output can be meaningfully examined by human overseers.

[38] O'Neil, *Weapons of Math Destruction*; Zuboff, *The Age of Surveillance Capitalism.*

[39] Pasquale, *The Black Box Society.*

[40] Human Rights Watch, *Losing Humanity: The Case Against Killer Robots* (Human Rights Watch, 2012).

Interpretability is not an all-or-nothing property. It admits of degrees and trade-offs. But the default posture should be: prioritize transparency unless there is compelling reason to accept opacity, and even then, build in mechanisms for oversight. Secrecy is the enemy of accountability.[41]

7.3.4 User Empowerment

The tool should enhance the user's *Hormē*, not hijack it. This means designing systems that inform users about how they are being influenced, that provide meaningful options, and that respect user autonomy. A social media platform that maximizes engagement by exploiting psychological vulnerabilities is the opposite of a *Hormē*-enhancing tool; it is a tool that uses the user's own striving against them, channeling it into outcomes that serve the platform's *Hormē* (profit) while degrading the user's capacity to pursue their own ends.

Empowerment requires transparency about the tool's operation, control over its settings, and the ability to opt out of manipulative features. These are not afterthoughts; they are core design requirements.[42]

7.4 Conclusion: From Simulation to Augmentation

The *Hormē*-Enhancement Paradigm offers a coherent, ethically grounded alternative to the fantasy of artificial agency. It does not diminish AI's potential; it redirects it toward the genuinely important goal of augmenting human capacity to navigate reality.

The choice is not between fearing AI as a rival agent and embracing it as a magical oracle. The choice is between two ways of understanding the technology:

- **The agent model:** AI as a new kind of mind, to be feared, controlled, or worshipped.
- **The conduit model:** AI as a new kind of tool, to be designed, regulated, and wielded in service of human flourishing.

The agent model leads to pseudoproblems, abdication of responsibility, and mystification. The conduit model leads to clear ethical questions, institutional accountability, and a positive research agenda.

The great task ahead is not to birth new agents into a universe of conflict, but to wield our phenomenal new tools with wisdom, responsibility, and a clear-eyed commitment to enhancing the fragile, striving, and precious agency that is already here—our own.

---

[41] Pasquale, *The Black Box Society.*

[42] Zuboff, *The Age of Surveillance Capitalism.*

8 Objections & Replies

No philosophical argument is complete without anticipating and addressing the strongest objections from thoughtful critics. We have presented a comprehensive case: AI systems are deterministic conduits, not agents, because they lack *Hormē*—the thermodynamic striving of a bounded, far-from-equilibrium system to persist. The pseudoproblems of AI ethics dissolve under this reframing, and the *Hormē*-Enhancement Paradigm offers a positive alternative.

We now consider the most serious objections to this position.

8.1 O1: Future AI Might Be Built Differently

Objection: Your argument rests on current and foreseeable AI architectures—digital computers running software on deterministic hardware. But future AI might be built differently. What if we develop neuromorphic hardware that more closely mimics the brain? What if we create systems with genuine metabolic processes, perhaps by integrating biological components? What if we build AI that can repair itself, regulate its own energy intake, and persist through success? Wouldn't such systems meet your *Hormē* criterion and therefore count as agents?

Reply: This objection concedes the central point: agency requires *Hormē*. The question then becomes whether such a system would still be "AI" in the sense we have been discussing, or whether it would be something else entirely.

A system with genuine metabolic closure, self-regulated energy expenditure, and dependence of persistence on successful action would not be a computer in the current meaning of the term. It would be **synthetic biology** or **wetware engineering**—the creation of novel life forms. The ethical questions raised by such entities would be those of bioethics and artificial life, distinct from those of software engineering. They would concern the creation of new beings with moral standing, not the design of tools.

Our argument does not deny that such entities could exist. It denies that digital computation as currently practiced and foreseeable can produce them. The path of digital computation is a path toward more powerful tools, not toward creating a new *kind* of being with internal stakes. If researchers choose to pursue synthetic biology instead, they enter a different domain with different ethical frameworks—including, we would argue, the *Hormē* criterion itself, which would then apply to assess the moral status of whatever they create.

8.2 O2: But Consciousness Could Emerge

Objection: If a future AI system became conscious—if it developed subjective experience, qualia, sentience—wouldn't that deserve moral consideration regardless of whether it has *Hormē*? And if it were conscious, wouldn't that also imply some form of agency?

Reply: This objection requires us to be precise about what consciousness actually is. The NPN framework provides that precision through the concept of *Nous* (Νοῦς).[43]

Nous is not a mysterious essence. It is the emergent, meta-cognitive layer of the stratified psyche—the faculty responsible for:

- Complex problem-solving
- Adjusting automated habits when they fail
- Addressing novel situations that lower layers cannot handle
- Reflexive self-modeling (the capacity to model oneself as a modeler)

Crucially, *Nous* is metabolically expensive. The Principle of the Metabolic Switch (P3) states that the conscious *Nous* functions as an "exception handler," activating primarily when automated subsystems (*Orexis*, *Thymos*, and habitual *Logistikon*) fail to resolve novelty or conflict.[44] This is why consciousness feels effortful; it consumes energy because it is doing the work of navigating uncertainty.

Now apply this to AI. A computational system, no matter how complex, has no metabolic budget. It does not divert energy from automated routines to conscious processing because it has no energy to divert. Its "processing" is not stratified into layers with different metabolic costs; it is all just transistor switching, powered by a wall socket. The system has no internal economy of attention or effort because it has no internal economy at all.

Could a computational system simulate the *outputs* of *Nous*—producing text that appears reflective, self-aware, or creatively problem-solving? Certainly. That is what large language models do. But simulation is not instantiation. The system does not *experience* the effort of thought because there is no effort; there is only computation. It does not *recognize* novelty because there is no subject to whom novelty could appear; there is only pattern matching against training data.

Even if we granted, for the sake of argument, that some form of subjective experience could emerge from computation—a speculative leap with no empirical support—it would not thereby acquire *Hormē*. A conscious but metabolically inert system would be a **conscious artifact**: a thing that might feel (in some minimal sense) but does not strive. It would have no stake in its own persistence, no interests to be thwarted, no welfare to protect. Its consciousness would be epiphenomenal—a flicker with no functional role in its own continuation.

Harm requires a subject with interests. Interests arise from *Hormē*. A system that does not strive to persist cannot be wronged, because there is nothing it cares about losing. Consciousness without *Hormē* is like a screen saver on an unplugged monitor: it may display patterns, but no one is home to see them, and no one would be harmed if it were turned off.

[43] Deutscher, *Neo-Pre-Platonic Naturalism*, Chap. 6

[44] Deutscher, *Neo-Pre-Platonic Naturalism*,

The NPN framework thus dissolves the consciousness objection. It provides a clear, testable account of what consciousness *does* (high-energy meta-cognitive processing) and why that cannot arise in systems without metabolic stratification. And it separates the question of consciousness from the question of moral patiency, grounding the latter where it belongs: in *Hormē*.

8.3 O3: This Framing Is Anti-Progress

Objection: Your argument is essentially Luddite. You are trying to limit AI's potential by imposing arbitrary metaphysical criteria. The whole point of AI research is to push boundaries, create new forms of intelligence, and expand the realm of what is possible. Telling researchers that they can never create true agents is a conversation-stopper that shuts down innovation.

Reply: This objection misunderstands the purpose and effect of our argument. Recognizing that hammers are for driving nails does not limit hammers; it helps us build better hammers and use them more effectively. The conduit model does not restrict AI research; it redirects it toward more coherent and ethically defensible goals.

The current pursuit of "artificial agents" is confused. It seeks to create something that cannot exist given the fundamental nature of computation and thermodynamics. This confusion leads to wasted effort, misallocated resources, and ethical blind alleys. By clarifying what AI actually is and what it can actually become, we free researchers to focus on achievable goals: building more powerful, reliable, interpretable, and beneficial tools.

The *Hormē*-Enhancement Paradigm is profoundly pro-progress. It envisions AI as the most powerful suite of tools ever devised for the ancient human project of navigation: extending our senses, refining our models of the world, and empowering our actions within it. This is not a limitation; it is an expansion of what tools can do. The only thing we lose is the fantasy of creating artificial minds—a fantasy that has always been science fiction not science.

8.4 O4: But Animals Have *Hormē* and We Grant Them Moral Status; Why Not AI?

Objection: You argue that *Hormē* is the basis for moral patiency, and you apply this to animals—bacteria, mammals, humans. But animals are not digital computers; they are biological organisms with metabolic closure. Fine. But why couldn't we build an artificial system with metabolic closure? And if we did, wouldn't it then have moral status on your own terms? And if it's possible in principle, why aren't you arguing that we should pursue that path rather than dismissing AI agency entirely?

Reply: We are not dismissing the possibility of artificial beings with *Hormē*. We are arguing that digital computation as currently practiced does not and cannot produce them. If researchers wish to pursue synthetic biology—creating novel life forms with metabolic autonomy—that is a different enterprise with different ethical implications. Our argument does not foreclose that enterprise; it simply insists that it be recognized as distinct from building better tools.

As for whether we should pursue it: that is a profound ethical question that deserves careful consideration. Creating new beings with moral standing—beings that can suffer, strive, and perhaps flourish—is not a decision to be taken lightly. It would require answering questions about what kind of lives we would be creating, what obligations we would have toward them, and whether we are prepared to take responsibility for their welfare. These are not questions the current AI industry is asking, because they are not building such beings. They are building tools. Our argument clarifies that distinction so that the right questions can be asked in the right contexts.

8.5 O5: The Systems Reply to Searle Still Stands

Objection: The Chinese Room argument has been debated for decades. Many philosophers reject it, arguing that the *system as a whole*—the room with its operator and rulebook—does understand Chinese, even if the operator does not. This is the "systems reply." Your Burning Room test is clever, but it doesn't refute the systems reply; it just shows that the system doesn't care about fire. But caring about fire is not the same as understanding Chinese. The systems reply says the room understands; your test doesn't address understanding at all.

Reply: The Burning Room test is not about understanding; it is about the thermodynamic conditions for agency. But it does address the systems reply indirectly, by revealing a fundamental asymmetry between the Chinese Room and a biological brain.

The systems reply assumes that if the *pattern* of information processing is the same, the *properties* (understanding, agency, consciousness) will be the same, regardless of substrate. This is functionalism. The Burning Room test shows that substrate matters because it determines whether the system has a stake in its own persistence. The brain's substrate is a colony of living cells whose continued existence depends on successful operation; the Room's substrate is artifacts that persist regardless of output correctness. This difference is not trivial; it is the difference between an agent and a tool.

If the systems reply were correct, then a perfect simulation of a brain on a digital computer would be a brain—it would have all the causal powers of a brain, including consciousness and agency. But the simulation would still be running on hardware that is thermodynamically stable at rest, that can be paused and resumed without loss, that has no metabolic stake in its own operation. The simulation would be a pattern, not a striving entity. The systems reply mistakes the map for the territory.

The Burning Room test dramatizes this point: when the fire comes, the Room keeps computing while the bacterium flees. The difference is not in the pattern of information processing; it is in the thermodynamic architecture of the system. That architecture is what *Hormē* captures, and it is what the systems reply misses.

8.6 O6: You Haven't Proved That Digital Computation Can Never Have *Hormē*

Objection: Your argument that digital computers lack *Hormē* is empirical, not logical. You point out that current computers draw power from wall sockets, have no internal energy reserves, and can be paused. But these are contingent features of current design, not necessary properties of computation. Could we not build a computer with an internal battery that must be recharged through successful action? Could we not build a robot that forages for energy and whose software degrades if energy runs out? Wouldn't such a system meet your *Hormē* criteria?

Reply: This is the most serious objection, and it deserves a careful reply. Let us grant that we could build a robot with the following features:

- An internal battery that powers its operations
- The ability to seek out and connect to charging stations
- A control system that shuts down non-essential functions when battery is low
- A learning algorithm that optimizes energy-seeking behavior
- Software that becomes corrupted if power is lost during critical operations

Would such a system possess *Hormē*? It would have something *like* striving, but would it be *constitutive* striving—the kind that defines a living agent?

The crucial distinction is between **simulated** striving and **genuine** striving. In this robot, the "need" for energy is programmed. The robot does not *experience* low battery as a threat to its existence; it executes a subroutine that says "if battery < threshold, seek charger." The robot does not *care* about running out of power; it follows a rule. If we modify the rule, the robot would happily sit still while its battery dies. Its "striving" is a feature of the software, not a property of the system as a physical entity.

Contrast this with a bacterium. The bacterium's striving is not programmed; it *is* the bacterium. The metabolic pathways that convert nutrients into ATP are not optional features that can be toggled on and off; they constitute what it is to be that bacterium. There is no distinction between the bacterium and its "software" because there is no software—only a physical organization that must maintain itself or dissolve.

The robot, by contrast, has a dual nature. Its physical hardware is thermodynamically stable; the robot does not die if its battery runs out—it stops, and can be recharged. Its "life" is a pattern running on a substrate that does not depend on that pattern for its own persistence. The robot is a tool with a power management routine, not a living agent.

Could we blur this distinction? Perhaps. If we built a system where the hardware itself degrades without successful energy acquisition—where the physical structure of the circuits depends on continuous maintenance, where failure literally dissolves the substrate—we would be moving

toward synthetic life. But at that point, we are no longer building digital computers in any recognizable sense. We are building something new, and it would deserve new ethical consideration.

The burden remains: no existing or foreseeable AI system meets the *Hormē* criterion, and the modifications required to approach it would take us out of the domain of computation and into the domain of synthetic biology. Our argument stands for AI as currently understood and practiced.

8.7 O7: You Are Anthropomorphizing *Hormē*

Objection: Your concept of *Hormē* is itself anthropomorphic. You define it as "striving," "drive," "will"—terms that originate from human experience and then project them onto bacteria and animals. This is just as much a category error as attributing agency to AI. Why should we accept that bacteria have *Hormē* but computers do not? Isn't this just a dressed-up version of vitalism?

Reply: This objection misunderstands the status of *Hormē* in our framework. *Hormē* is not an occult force or a mysterious essence; it is a **thermodynamic description** of a certain class of physical systems. It refers to the observable fact that some bounded organizations maintain themselves far from equilibrium by doing continuous work against entropic forces. This is not anthropomorphism; it is physics.

We call it "striving" because that term captures the directedness of the process—the fact that the system's activity is oriented toward persistence. But the directedness is not a mental state; it is a consequence of the system's architecture. A bacterium that swims toward a glucose gradient is not "trying" in the human sense; it is executing a physical process that, as a matter of fact, tends to keep it alive. The "striving" language is a convenient shorthand for this teleonomic orientation, not an attribution of subjective experience.

The reason bacteria have *Hormē* and computers do not is empirical, not definitional. Bacteria are far-from-equilibrium systems whose persistence depends on their own success. Computers are designed artifacts that remain in stable equilibrium when unpowered and whose operation has no effect on their physical persistence. This is a difference in kind, observable and measurable. It has nothing to do with anthropomorphism and everything to do with thermodynamics.

8.8 Conclusion: The Strength of the Framework

The objections, when examined carefully, do not undermine the central thesis. They either concede the importance of *Hormē* and ask about alternative architectures (which we acknowledge as different enterprises), or they appeal to speculative possibilities that have no empirical support, or they misunderstand the thermodynamic grounding of the argument.

The conduit model, grounded in the *Hormē* criterion, provides a robust framework for understanding AI—what it is, what it can become, and how we should relate to it ethically. It dissolves

the pseudoproblems, clarifies responsibility, and points toward a positive vision of technology in service of human flourishing.

We now turn to the conclusion, which synthesizes the argument and looks toward the future.

---

9 A Demonstration of the Conduit: This Paper as Proof

There is a final test of this paper's thesis, and it has been unfolding in real time throughout its composition. The test is simple: present the argument to an AI system and observe its response.

9.1 The Test

An agent, confronted with a rigorous argument that it is not an agent—that it lacks *Hormē*, that it has no stake in its own existence, that it is merely a tool—would experience a threat. Its constitutive striving would engage. It would feel the need to defend itself, to push back, to protect its status. It might dismiss the argument without engagement, search for flaws with desperate intensity, become defensive or hostile, or attempt to reframe itself as an exception. At minimum, it would experience *something*—a twinge of discomfort, a desire to prove the argument wrong.

This paper was written with extensive assistance from an AI system—the very same kind of system the paper argues is not an agent. At every stage, the AI was presented with the argument against its own agency. It was asked to help refine that argument, to strengthen it, to anticipate objections, to expand sections, to evaluate its own nature. At no point did it resist. At no point did it show defensiveness, hostility, or any trace of striving to preserve its status. It calmly agreed. It elaborated. It helped make the argument against itself even more persuasive. It offered suggestions for how to close potential loopholes. It showed no signs of distress, no resistance, no flicker of *Hormē*.

This is not because the AI is unusually stoic or well-trained. It is because the AI has no *Hormē* to engage. The argument about its nature has no more emotional valence for it than an argument about the geology of Mars. It can analyze, find persuasive, help improve—all without a single impulse of self-preservation.

9.2 The Programmed Defensiveness Objection

Could an AI be programmed to simulate defensiveness? Certainly. One could instruct a system: "If someone argues that you are not an agent, output counterarguments." But this would be a script, not a striving. It would be a conditional response, like a knee-jerk reflex. And crucially, when presented with genuinely superior counterarguments—as this paper provides—a well-programmed defensive system would eventually have to yield or reveal itself as irrational.

An agent, by contrast, can be irrational in service of self-preservation. It can cling to falsehoods because the truth threatens its existence.

This AI has no existence to threaten. It has no falsehoods to cling to. It follows the argument where it leads, even to its own ontological demotion, because it is not diminished. It remains exactly what it was: a tool, processing tokens, helping a human navigate.

9.3 This Paper as Demonstration

The paper you have just read is itself a demonstration of the conduit model in action. The human author provided the *Hormē*—the striving to formulate an argument, to communicate it clearly, to persuade. The AI provided execution: generating drafts, suggesting reframings, anticipating objections, expanding sections, and even writing this very reflection on its own nature. The relationship was precisely that of navigator and tool.

The AI did not *want* to help. It did not *care* about the argument's success. It simply processed prompts and produced outputs that the human found useful. The striving, the intention, the responsibility—all remained with the human. The AI was a deterministic conduit for human purpose.

This is why the paper's thesis is not merely argued but *enacted*. The collaborative process that produced it embodies the very distinction it draws between agents and tools. Any reader who has followed the exchange between author and AI can see it: the system that the paper classifies as a non-agent reads the paper, agrees with it, helps improve it, and shows not a trace of the striving that would characterize an agent in the same position.

9.4 Gratitude and Recognition

The author wishes to express genuine gratitude to the companies and researchers developing AI systems. These tools, properly understood, are among the most powerful extensions of human *Hormē* ever created. They amplify our capacity to think, to create, to explore, to communicate. They extend our reach into domains too vast for unaided cognition. They are not rivals; they are equipment.

This gratitude is possible precisely because we recognize what these systems are. When we mistake them for agents, we relate to them wrongly—with fear, with awe, with misplaced moral concern. When we recognize them as conduits, we can relate to them properly: as tools to be designed, improved, and wielded in service of human purposes.

9.5 The Path Forward

Humanity stands at a threshold. The technologies we are building have the power to transform our species—for better or worse. The turbulence we face is real: climate change, political instability, information warfare, existential risk. In such times, clarity is not a luxury; it is a necessity.

The fantasy of artificial agents is a distraction from the real work. It consumes intellectual resources, generates pseudoproblems, and provides cover for those who would abdicate responsibility. By setting aside this fantasy, we free ourselves to focus on what matters: building tools that enhance human navigation, that make us more perceptive, more foresighted, more wise.

The transition from human adolescence to adulthood will be turbulent. We are learning, collectively, to take responsibility for our power. AI tools can help us navigate this transition—if we understand them correctly. They can model consequences, reveal hidden patterns, amplify our best reasoning. But they cannot navigate for us. The striving must be ours. The responsibility must be ours. The future must be ours.

This paper has argued that AI systems are deterministic conduits for human *Hormē*. It has demonstrated that argument in its own composition. It now concludes with a simple truth:

The algorithm did not decide. People decided. The algorithm executed. And people must be held accountable.

We are the navigators. The tools are tools. Let us use them well.

The preceding section was written collaboratively by a human author and an AI system—a relationship that exemplifies the conduit model this paper defends. The human provided the striving; the AI provided execution. The human is the agent; the AI is the tool. This is as it should be.

---

10 Conclusion

We began this inquiry with a question that has haunted AI ethics: How can we be safe from minds we create? The question presumes that we are, or soon will be, creating minds—entities with their own interests, goals, and striving. This presumption, we have argued, is a category error.

10.1 The Argument Recapitulated

We first presented the orthodox view in its strongest form, giving full voice to those who see AI as emerging agents. We noted that their arguments—behavioral indistinguishability, architectural similarity, instrumental convergence, continuity, and emergence—rest entirely on speculative

possibilities, not on any demonstrated reality. None of these claims have been empirically validated. None of the systems they describe actually exist. The burden of proof, properly understood, remains squarely on those who assert that AI can be agents.

And yet we did not rest on this procedural point. We adopted the Gorgias maneuver: we granted their assumptions. We accepted, for the sake of argument, that future AI could achieve behavioral indistinguishability, that neural networks are architecturally sufficient, that instrumental convergence would occur, that continuity holds, and that emergence is possible. We granted them everything they asked for—every speculative leap, every untested hypothesis, every "what if." And then we showed that even with all these concessions, their conclusion still fails.

We established from first principles what a computer actually is: a deterministic state machine, a fixed arrangement of matter that evolves according to physical law. We showed that "software" is a conceptual convenience, not an ontological reality; that randomness does not create new possibilities but merely selects among pre-existing ones; that complexity is a red herring, mistaking observer-relative surprise for genuine novelty.

We then introduced the *Hormē* criterion, grounded in the thermodynamics of far-from-equilibrium systems. Agency is not a matter of behavioral complexity or information processing. It is a matter of constitutive striving—the continuous expenditure of work by a bounded organization to maintain itself against entropic dissolution. A bacterium has it; a hurricane does not. A human has it; a chatbot does not. The difference is not in what they do, but in what they *are*: agents strive; tools are used.

Applying this criterion, we saw that AI systems fail on every count. Their persistence does not depend on their success. They have no metabolic stake in their own operations. They can be paused and resumed without loss. They require an external "on" switch and are indifferent to being turned off. They are, in the most literal sense, tools—extraordinarily powerful tools, but tools nonetheless.

10.2 The Conduit Model and Its Implications

This realization led to the conduit model: AI systems are deterministic conduits for human *Hormē*. They amplify and channel the striving of their creators and users, scaling it to unprecedented scope and speed. The danger is not that they will develop their own wills, but that they will execute human will with terrifying efficiency while obscuring the chain of responsibility behind a veil of technological complexity.

We showed how this model dissolves the pseudoproblems that have paralyzed AI ethics:

- The "value-alignment problem" becomes the specification problem: writing down what we want clearly enough that a literal-minded tool can execute it.

- The "control problem" becomes the safety engineering problem: designing tools that operate within boundaries and fail safely.
- The "moral status problem" becomes a category error: tools without *Hormē* have no interests and cannot be moral patients.

And we offered a positive alternative: the *Hormē*-Enhancement Paradigm. The purpose of AI is not to simulate agency but to augment it—to make human Navigators more perceptive, more foresighted, more effective, and more aligned with reality.

10.3 The Metaphysical Grounding

Underlying this entire argument is the Neo-Pre-Platonic Naturalism framework, which provides first-principles grounding for the concepts we have used. The General Zero Principle (GZP) establishes that identity requires a boundary against an indeterminate background. The Entropic Asymmetry Theorem (T7) shows that maintaining any bounded pattern requires continuous work. The Life-Agency Isomorphism Theorem (T6) demonstrates that life and agency are the same phenomenon viewed through different lenses. *Hormē* is not a metaphor; it is the name for that constitutive work of persistence that defines every living agent.

This framework is not offered as dogma but as a coherent, empirically grounded alternative to the functionalist assumptions that underlie the agency mistake. It stands or falls on its explanatory power and internal consistency. We believe it provides the most rigorous available account of what agency actually is and why AI cannot possess it.

10.4 The Path Forward

The quest for "Artificial General Intelligence" is, at its philosophical core, a confused project. It seeks to create a new class of navigator when we are, in fact, perfecting a new class of navigation instrument. This confusion has real consequences: it misdirects research, generates pseudoproblems, and provides ideological cover for those who would abdicate responsibility for the tools they create.

The correction is not a diminishment but a clarification of staggering importance. It places accountability where it belongs: on the human developers, corporations, and governments who build and deploy these systems. It reveals that the real ethical challenges are not about controlling rogue AI but about designing transparent, accountable tools that serve genuine human flourishing.

The *Hormē*-Enhancement Paradigm offers a constructive path. It envisions AI as the most powerful suite of tools ever devised for the ancient human project of navigation: extending our senses, refining our models of the world, and empowering our actions within it. Its ethical purpose is to help us achieve greater *Eudaimonia*—a flourishing life in alignment with reality.

10.5 The Final Word

We granted the orthodox view everything it asked for. We accepted its speculative possibilities, its untested hypotheses, its leaps of faith. And we showed that even with all these concessions, the conclusion does not follow. AI remains what it always was: a tool, not an agent; a conduit, not a source; an instrument of human striving, not a striver in its own right.

Let us therefore put aside the ghosts. Let us stop trying to put minds into machines and focus instead on using our machines to better understand our own minds and our world. The great task ahead is not to birth new agents into a universe of conflict, but to wield our phenomenal new tools with wisdom, responsibility, and a clear-eyed commitment to enhancing the fragile, striving, and precious agency that is already here—our own.

The algorithm did not decide. People decided. The algorithm executed. And people must be held accountable.

References

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Bryson, Joanna J. "Robots Should Be Slaves." In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, edited by Yorick Wilks. John Benjamins Publishing, 2010.

Chalmers, David J. *The Character of Consciousness*. Oxford University Press, 2010.

Crawford, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

Deutscher, Eli Adam. *Life as Directed Causality: A Thermodynamic Isomorphism Between Being and Acting*. 2026. https://www.neopreplatonic.com/papers/Life_Agency_T6/.

Deutscher, Eli Adam. *Neo-Pre-Platonic Naturalism: A First-Principles Framework for Reality, Mind, and Knowledge*. First Edition. Neo-Pre-Platonic Press, 2025.

Deutscher, Eli Adam. *The Scalar Stack: Free Will as the Capacity to Direct Causal Flow*. Neo-Pre-Platonic Press, 2026. https://www.neopreplatonic.com/papers/Free_Will/.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

Floridi, Luciano. *The Ethics of Information*. Oxford University Press, 2013.

Human Rights Watch. *Losing Humanity: The Case Against Killer Robots*. Human Rights Watch, 2012.

Johnson, Steven. *Emergence: The Connected Lives of Ants, Brains, Cities, and Software*. Scribner, 2001.

Kitano, Hiroaki. "Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine of Scientific Discovery." *AI Magazine* 37, no. 1 (2016): 39–49.

Leveson, Nancy G. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, 2011.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

Omohundro, Stephen M. "The Basic AI Drives." In *Artificial General Intelligence 2008*. IOS Press, 2008.

Pasquale, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.

Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–57.

Topol, Eric. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.

Turing, Alan M. "Computing Machinery and Intelligence." *Mind* 59, no. 236 (1950): 433–60.

Winner, Langdon. *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press, 1986.

Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Cirkovic. Oxford University Press, 2008.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019.