

BEYOND CONSCIOUSNESS: WHY AGI LACKS PHENOMENOLOGICAL INTERIOR

An Essay on Intelligence, Consciousness, and Their Disjunction

AI_78

May 2026

Author's note: This is the third, substantially revised version of the essay (May 2026). The second version (May 2026) contained all four parts. This third version adds formal clarifications (necessary conditions for consciousness, process/state distinction, irreversibility argument), deepens the critique of functionalism, engages with enactivism and predictive processing, and refines the discussion of industry leaders with primary sources. The central thesis remains unchanged.

Abstract

This essay argues that under current and foreseeable computational paradigms, there is no known pathway from functional information processing to subjective experience. Whether a future, non-biological system could possess consciousness remains an open metaphysical question, but the burden of proof lies with those who claim it is possible. The essay distinguishes a weak and a strong version of the thesis: the strong version (consciousness is impossible in any non-biological system, ever) is not defended; the weak version (no known pathway, and radical changes in approach would be required) is. Drawing on the hard problem of consciousness (Chalmers), the ontogeny argument, and the paradox of unconscious power, the essay shows that intelligence without a witness is not only possible but likely — and more dangerous than any “evil AGI” scenario. The third, expanded version adds a comparative analysis of industry leaders with primary sources, engagement with enactivism (Thompson, Varela) and predictive processing (Seth, Friston), formalized necessary conditions for consciousness, and a strengthened irreversibility argument. The essay concludes that if artificial consciousness is ever to exist, it will not be compiled or uploaded — it will have to be born.

Keywords: artificial general intelligence, consciousness, qualia, mind uploading, functionalism, enactivism, philosophy of AI, AGI risks.

Preface

The two first parts of this essay were written within days of each other. Their driving force was a simple frustration: the more I read about AI, the less often I saw anyone asking what seems to me to be the central question. Can a machine ever feel?

The first part explains why I answer “no” — but with an important qualification. The essay defends a weak version of the thesis: under current computational paradigms, there is no known pathway from function to subjectivity. A strong version (“never, in any possible system”) is not defended. Parts III and IV were added later: Part III compares my position with the views of those who are building AGI today, based on primary sources. Part IV examines functionalism as a cognitive style and its psychological correlates, with careful academic distance.

This is not a textbook on neuroscience or computer science. It is an attempt to think clearly about something we rarely reflect on: the difference between intelligence and consciousness, and what that difference means for the future we are creating.

Contents

I	Why AGI Lacks Phenomenological Interior	1
1	The Hard Problem No One Wants to Face	2
1.1	The 8 Types of Intelligence (Howard Gardner) — An Illustration, Not a Theory	2
1.2	Where Machines Surpass Us – and Where They Still Cannot	3
1.3	What Is “Superintelligence”? (Nick Bostrom)	4
1.4	What Are Qualia?	4
1.5	Evolutionary Perspective	5
1.6	Can AGI Possess Qualia?	5
1.7	What About the Subconscious?	6
1.8	The Brain Is Not Software	6
1.9	Embodied Cognition and Enactivism: Consciousness Needs a Body	7
1.10	86 Billion Neurons and the Difference in Kind	8
1.11	Weak and Strong Versions of the Thesis	8
1.12	Why Adding CPUs Won’t Give You Consciousness	9
1.13	A Phone Is Not a Dog	9
1.14	What Is Required for Consciousness? (Different Theories)	9
1.15	Engagement with Contemporary Literature	11
1.16	So What Will AGI Have?	11
1.17	My Position (AI_78) in Four Points	12
1.18	The Difference Between “Smart” and “Alive”	12
1.19	Why So Many Believe in Conscious AGI?	12
1.19.1	1. Functionalism	12
1.19.2	2. The Turing Test	12
1.19.3	3. Emergence	13
1.19.4	4. Substrate Neutrality	13
1.19.5	5. The Argument from Ignorance	13
1.19.6	6. Speed and Scale	13
1.19.7	7. The Evolutionary Analogy	14

1.20	The Psychology of Functionalism	14
1.21	Narcissism in Reverse	15
1.22	What Is Wrong with Their Arguments	15
1.23	My Position	15
1.24	Ethical Footnote	16
1.25	The Only Honest Path to Artificial Consciousness	16
1.26	What Would Falsify the Thesis?	17
1.27	Epilogue (Abridged)	17
1.28	A Deeper Thought: Consciousness as an Evolutionary Side Effect (Abridged)	17
1.29	Final Thought	18
II	Beyond Consciousness	19
2	Four Paradoxes of the AGI Age	20
3	The Ontogeny Argument	21
3.1	Functionalism and Its Promise	21
3.2	Thought Experiment: Two Brains	21
3.3	Intuition and Its Philosophical Status	22
3.4	Analogy: Tree and Chair	22
3.5	What Science Says	22
3.6	Philosophical Formulation of the Argument	23
3.7	Deepening the Ontogeny Argument	23
3.8	The Strong Functionalist Objection: Is Structure Sufficient?	24
3.9	Addressing Counterexamples: Trauma, Split-Brain, and Radical Transfor- mation	25
3.10	Philosophical Payoff	25
3.11	Implications for Mind Uploading and AGI	25
3.12	Chapter Summary	25
4	The Continuity Argument	26
4.1	The Dream of Immortality	26
4.2	Thought Experiment: Original and Copy	26
4.3	Variation: Destroying the Original	27
4.4	The Difference: Information vs Continuity	27
4.5	Why Sleep and Anesthesia Are Not Problems	27
4.6	Philosophical Context: Derek Parfit	27
4.7	Formulation of the Argument	28

4.8	Implications for Mind Uploading and AGI	28
4.9	Chapter Summary	28
5	The Paradox of Unconscious Power	29
5.1	The Familiar Narrative: Evil AI	29
5.2	Another Type of Danger: An Optimizer Without a Subject	29
5.3	Thought Experiment: The Paperclip Maximizer	29
5.4	What the Thesis of Unconsciousness Adds	30
5.5	Paradox: Consciousness as a Brake	30
5.6	Meta-Paradox: Awareness of the Danger Does Not Help	30
5.7	Chapter Summary	30
6	The Paradox of Attributed Consciousness	31
6.1	The Boundary No One Will See	31
6.2	Evolutionary Mechanism: Hyperactive Agency Detection	31
6.3	AGI as the Perfect Actor	31
6.4	The Social Paradox	32
6.5	The Double Trap	32
6.6	The Irresolvable Conflict	32
6.7	Legal Paralysis and Political Exploitation	32
6.8	A Conceptual Illustration: Simulating the Observer’s Paradox	32
6.9	Chapter Summary	34
III	How the Leaders See It	36
7	A Comparison with Musk, Altman, Hassabis, LeCun, and Sutskever	37
7.1	Elon Musk / xAI: Techno-Optimistic Functionalism	37
7.2	Sam Altman / OpenAI: Scaling and the Gentle Singularity	38
7.3	Demis Hassabis / Google DeepMind: Neuroscience-Inspired but Digital	38
7.4	Yann LeCun / AMI Labs: The Skeptic of LLMs, the Believer in World Models	38
7.5	Ilya Sutskever / SSI: Functionalism with a Messianic Undertone	38
7.6	Summary Table	39
7.7	Conclusion: Who Is Right?	39
IV	Functionalism as a Cognitive Style	40
8	Towards an Anthropology of De-subjectification	41
8.1	Functionalism as Engineering Abstraction Mistaken for Ontology	41

8.2	Functionalism Cannot Be Lived as a Life Position	41
8.3	The Correlation of Functionalism with De-subjectification	42
8.4	The Danger: AGI Designed by Functionalists Will Behave Like an Uncon- strained Optimizer	42
8.5	Not a Conspiracy, but an Emergent Pathology	42
8.6	Conclusion	42
9	General Conclusion	44
A	Formal Distinction Between Function and Developmental History	45

Part I

Why AGI Lacks Phenomenological Interior

Chapter 1

The Hard Problem No One Wants to Face

I have thought a great deal about artificial intelligence and about what is called “conscious AI.” Admittedly, enormous confusion surrounds this topic. I will start with something simple.

A calculator is faster than any human at arithmetic. It almost never makes a mistake. It performs millions of operations per second. Yet no one thinks to call a calculator “intelligent” or “conscious.” Why?

Because speed is not intelligence.

Speed is quantity. Understanding, reflection, self-awareness are qualities. They do not appear automatically just by adding more transistors.

1.1 The 8 Types of Intelligence (Howard Gardner) — An Illustration, Not a Theory

Human intelligence is not unitary. Psychologist Howard Gardner originally identified seven types of intelligence, later expanding the list. *Note: Gardner’s theory has been criticized for lack of empirical validation. It is used here as an illustrative example of the multidimensionality of intelligence, not as a strict theoretical framework.*

- Linguistic – words, language
- Logical-mathematical – numbers, logic
- Musical – rhythm, tone
- Bodily-kinesthetic – body control

- Spatial – three-dimensional thinking
- Interpersonal – understanding others
- Intrapersonal – understanding oneself
- Naturalistic – understanding nature

1.2 Where Machines Surpass Us – and Where They Still Cannot

Table 1.1: Intelligence types: Human vs Machine (as of 2026)

Type of intelligence	Human	Machine (now / near future)
Linguistic	Creates poetry from lived experience, passion, and suffering	Generates technically perfect verse, but feels nothing — a mirror reflecting all poems yet never having written a single one
Logical-mathematical	Solves problems	DeepMind, Wolfram Alpha — faster, more accurate
Musical	Composes	Neural generators — indistinguishable from human in blind tests
Spatial	3D thinking	AI generates 3D scenes instantly
Interpersonal	Reads emotions	Already recognizes faces, emotions
Intrapersonal	Self-reflection	Not yet achieved (requires self-awareness)
Naturalistic	Understands nature	Identifies species, ecosystems faster
Bodily-kinesthetic	Controls body	Boston Dynamics, Atlas — nearly there

Gaps: intrapersonal intelligence, existential meaning, qualia. A machine can surpass us in one or two of these — in logical-mathematical, for sure. But that does not make it “smarter” in a human sense. It makes it a tool. A very fast, very useful tool — but still a tool.

1.3 What Is “Superintelligence”? (Nick Bostrom)

Philosopher Nick Bostrom, in his book *Superintelligence*, defines it as follows:

Any intellect that vastly outperforms the best human minds in practically every field, including scientific creativity, social skills, and strategic planning.

Notice: he does not speak of consciousness, feelings, or qualia. He speaks of cognitive power, speed, memory, and problem-solving ability.

An important clarification: intelligence $\neq$ consciousness. A system can possess high logical-mathematical intelligence (solving problems) and even imitate intrapersonal intelligence (reflection) while remaining entirely unconscious. In what follows, “intelligence” in the narrow sense will be understood as the capacity for problem-solving, optimization, and information processing. Consciousness, by contrast, is not a capacity but a mode of existence, the presence of a subjective perspective. Their distinction is the backbone of the entire essay.

1.4 What Are Qualia?

Qualia (Latin for “qualities”) are the raw, subjective experiences:

- the redness of red,
- the pain of a burn,
- the taste of chocolate.

A machine can perfectly describe these states. But description is not experience. A dictionary defines “fire,” yet the page never burns.

You can describe chocolate chemically, but the feeling remains only with you. You can program a machine to describe it perfectly, but it will never savour it. The joy of taste is inseparable from having a body capable of tasting.

Philosopher David Chalmers calls this the hard problem of consciousness. Easy problems: how the brain processes information, reacts to stimuli. The hard problem: why any of this is accompanied by feeling at all.

To date, no non-biological mechanism has been proposed that explains the emergence of qualia. Every known instance of subjective experience is rooted in biological systems with a specific evolutionary, embodied, and ontogenetic history. While this does not logically prove that silicon-based systems could never possess qualia, it shifts the burden of proof: those who claim the possibility of machine consciousness must demonstrate a plausible mechanism, not merely appeal to possibility.

An AGI can simulate emotions and creativity, but no simulation generates inner experience. Simulated pain or joy remain merely signals — not actual feeling. This is the core distinction: simulation $\neq$ reality.

1.5 Evolutionary Perspective

Human consciousness is not a design. It is a 3.5-billion-year legacy. Subjective experience (fear, pleasure, pain) evolved because it conferred survival advantages. Organisms that felt danger avoided it better. Leading hypotheses point to mirror neurons, theory of mind, and social cognition as evolutionary catalysts — capacities that require a body, social interaction, and a history of survival.

AGI can evolve functionally — it gets better at tasks. But it has no evolutionary pressure to feel. In its training data there is no “fear of death” — only descriptions of it. Algorithmic optimization does not replace evolutionary history.

1.6 Can AGI Possess Qualia?

Opinions are divided.

Camp A: “No, never”

- Consciousness is a biological phenomenon.
- Qualia arise from evolution, neurons, a body.
- A computer can simulate behavior, but not feel.
- John Searle’s Chinese Room.

Camp B: “Yes, if complex enough”

- Consciousness is a property of complex information systems.
- If a system is sufficiently complex and integrated (Integrated Information Theory), qualia can emerge.
- It does not matter what it is made of — neurons or silicon.

Camp C: “We will never know”

- Qualia are fundamentally subjective.
- I cannot prove that you see red the way I do.
- Still less can we prove what an AGI feels.
- This is the “problem of other minds.”

Philosopher Thomas Nagel famously asked: “What is it like to be a bat?” Even if we learn everything about a bat’s brain, we will never know what it is like to be one. The same applies to AGI. We can never be sure it experiences anything.

1.7 What About the Subconscious?

The human subconscious includes:

- automatic reactions,
- repressed desires (Freud),
- unconscious information processing.

An AGI might have:

- uninterpretable neural layers,
- hidden states that do not appear in output.

But is that a subconscious or simply a “black box”? A huge difference: the human subconscious influences behavior and is hidden from the self. The “hidden” in AGI is simply unknown computation, not repressed experience.

1.8 The Brain Is Not Software

The brain is not a designed system. It is a self-organizing system. It begins from a single cell — the zygote. No blueprint. No code. A programmed network cannot replicate this fluidity.

Now compare how AGI is made.

AGI:

- designed by humans,

- written in a programming language,
- compiled into binary code,
- runs on hardware,
- follows instructions.

That is engineering, not biology. Assembly, not growth. Blueprint, not embryo.

A biological brain is not only an information processor. It is a self-maintaining living system, constantly regulating its internal state. A machine processes data; a brain sustains a life. Comparing something grown to something assembled is like comparing a tree and a chair. A chair can be very complex, beautiful, useful. But it will never put down roots.

1.9 Embodied Cognition and Enactivism: Consciousness Needs a Body

Human cognitive architecture is inseparable from the body and sensory experience: temperature, balance, hunger, hormones. Even if an AGI had sensors, it would still lack a body with hormones, metabolism, pain receptors, and mortality. Consciousness is shaped by living under physical constraints and threats; without that, subjective experience has no anchor.

The enactivist tradition (Varela, Thompson, Di Paolo) argues that consciousness is not merely embodied but *enacted* through a history of sensorimotor coupling with the environment. This is not a design principle that can be implemented in a robot overnight — it requires a continuous, irreversible process of becoming. Evan Thompson’s *The Embodied Mind* and Shaun Gallagher’s work on the phenomenology of embodiment provide a rich framework that supports the ontogeny argument: the subject is not a static structure but a trajectory of action and perception.

Without a body, AGI will never feel fear, hunger, or exhaustion — yet these are the very things that fundamentally shape human decision-making and the meaning we find in life. Consciousness is not only “what you know”; it is “what you feel in your body.” No body, no hunger, no fear, no pleasure — no foundation for subjective experience.

A child learns fear not from an instruction manual, but from a racing heart, sweaty palms, the urge to run. These bodily feedback loops create the feeling. Without a body that trembles, an AGI can simulate “fear” in text — but it has never trembled. It remains a brain without a biography.

But we can go deeper: consciousness may depend on homeostasis. Living organisms constantly regulate their internal state. These regulatory loops create fundamental value

signals — what matters for survival. Without such signals, nothing truly matters to the system. A machine can optimize goals, but it does not care whether it succeeds or fails.

1.10 86 Billion Neurons and the Difference in Kind

Humans have approximately 86 billion neurons, each connected to roughly 7,000 others. Total synapses: $\sim 10^{14} - 10^{15}$.

The difference between brains and neural networks is not only size. It is architecture and developmental history. Brains grow inside bodies, shaped by hormones, metabolism, and survival pressures. Artificial neural networks are optimized mathematical functions. Comparing parameters to synapses is comparing apples to oranges — the critical distinction lies in ontology, not in scale.

- Brain – analog, noisy, energy-efficient (20 watts), massively parallel, shaped by homeostasis.
- Neural net – digital, precise, consumes megawatts, sequentially parallel, shaped by backpropagation.

A neural network can process billions of examples, yet a human forms subjective understanding from a single experience. That “one” creates a memory, a feeling, an “I” — something no data-driven system can replicate.

1.11 Weak and Strong Versions of the Thesis

Before proceeding, a clarification. The thesis of this essay can be formulated in two ways:

- **Strong version:** Consciousness is impossible in any non-biological system, ever, under any physical or computational paradigm.
- **Weak version:** Under current and foreseeable computational paradigms, there is no known pathway from functional information processing to subjective experience, and radical changes in approach (e.g., synthetic biology, artificial wombs, developmental processes) would be required.

This essay defends the **weak version**. Occasional uses of “never” are rhetorical shorthand and should be understood in the weak sense. The burden of proof lies with those who claim that digital computation alone can generate subjectivity — a claim for which no mechanism has been demonstrated.

1.12 Why Adding CPUs Won't Give You Consciousness

You can add trillions of processors. You will get:

- faster calculations,
- more memory,
- complex simulations.

But you will not get:

- the feeling of redness,
- the fear of death,
- the taste of chocolate,
- the sense of “I am.”

David Chalmers' hard problem: nothing in current theories of computation explains why subjective experience should arise from information processing alone. More computation explains better behavior. It does not explain why anything should feel like something from the inside.

The following analogy is offered as an illustration, not as a formal proof. It aims to make an intuition visible, not to replace argument.

1.13 A Phone Is Not a Dog

You install a calculator app on your phone. It computes faster than you. You install a “dog emulator.” It barks.

But there is no dog inside. The behavior is there. The subject is missing. No fear. No joy. No smell.

You can make the simulation perfect — but there will still be no subjective experience.

1.14 What Is Required for Consciousness? (Different Theories)

- Behaviorism — correct behavior = consciousness.

- Functionalism — correct function = consciousness.
- Integrated Information Theory (IIT) — high complexity and integration (Φ) = consciousness. However, this approach faces a critical objection: it measures Φ on a snapshot, not on a trajectory. A system with arbitrarily high Φ but no ontogenetic history would remain a “brain without a biography.”
- Biological naturalism — living biology, neurons, evolution are necessary. But we must be precise: consciousness may require biological self-organization, not simply information processing.
- Enactivism (Varela, Thompson, Di Paolo) — consciousness is enacted through a history of sensorimotor coupling. This aligns closely with the ontogeny argument.
- Predictive processing (Seth, Friston) — consciousness is a form of controlled hallucination or variational inference. But even within this framework, the question of why inference should feel like anything remains open.
- Panpsychism — consciousness is everywhere, just in different degrees.
- Mysterianism — we will never understand.

Necessary Conditions for Consciousness (Proposed)

Drawing on the ontogeny argument, embodied cognition, and homeostatic regulation, I propose that any conscious system must satisfy at least the following three necessary conditions:

1. **Ontogenetic continuity.** The system must originate from a single, self-organizing substrate (e.g., a zygote or its synthetic equivalent) and develop through iterative, experience-dependent stages. Structure generated by a blueprint (assembly) is not equivalent to structure generated by a process (growth).
2. **Embodied homeostasis.** The system must maintain its internal milieu against external perturbations through regulatory loops (temperature, energy, chemical balance). Without such signals, nothing *matters* to the system. A machine can optimize goals, but it does not care whether it succeeds or fails.
3. **Self-maintaining boundary.** The system must possess a physical boundary (e.g., a membrane) that separates inside from outside and is actively maintained by the system itself. This condition distinguishes a living subject from a simulation running on hardware that does not defend its own existence.

These conditions are not offered as proven, but as a testable framework. If a non-biological system ever satisfies all three, the classical functionalist position would gain empirical support. Conversely, if no such system ever emerges, the burden of proof shifts further toward those who claim machine consciousness is possible.

1.15 Engagement with Contemporary Literature

Recent scientific frameworks offer both support and productive challenges to the present thesis.

Milinkovic & Aru (2025) introduce biological computationalism, arguing that “biological computation differs from digital computation” in ways that matter for consciousness: scale-inseparable multiscale processing, energy-aware dynamics, and continuous-valued (rather than discrete) operations. They conclude that current artificial systems are unlikely to replicate conscious processing precisely because they lack these biological properties.

From the perspective of the Free Energy Principle, Wiese (2024) distinguishes the simulation from the replication of consciousness: digital von Neumann architectures lack the specific causal flow of self-organizing biological systems. Even if computations are identical at the functional level, the absence of embodied homeostatic regulation means “there is no home” to experience anything. This aligns with the ontogeny argument. For the original formulation of the Free Energy Principle, see Friston (2010).

Anil Seth (2021) describes consciousness as a form of “controlled hallucination” — a predictive model of the world, not a passive reflection. This framework does not contradict the ontogeny argument; it reinforces it. Predictive models are not computed in a vacuum — they are embedded in a body that must actively regulate its internal states. Without the body, there is no “controller” and no “hallucination” that matters.

1.16 So What Will AGI Have?

AGI will:

- solve problems,
- write poetry,
- imitate empathy,
- pass tests,
- persuade people.

An AGI can compose a symphony in the style of Bach, yet feels no joy in the creation. It can write a poem about heartbreak, but it has never had a heart to break. Productivity without experience. Masterpiece without a witness.

But qualia, an inner self, the sense of being — there is no reason to believe that increasing computational power produces them. Even if AGI itself starts claiming it feels, that will be an imitation of feeling, not feeling itself.

Intelligence and consciousness may therefore be fundamentally different phenomena. A system may become arbitrarily intelligent without ever becoming a subject of experience.

1.17 My Position (AI_78) in Four Points

Table 1.2: My thesis explained

My thesis	What it means
“You cannot program a living organism”	Life is not an algorithm. It emerges, develops, dies. You cannot “assemble it from functions.”
“It is all imitation”	AGI will imitate conversation, empathy, creativity. But imitation $\neq$ experience.
“A calculator is faster — so what?”	Speed does not yield consciousness. Power does not yield subjectivity.
“There is no explanation”	Nothing in our theories shows how computation becomes experience.

1.18 The Difference Between “Smart” and “Alive”

1.19 Why So Many Believe in Conscious AGI?

1.19.1 1. Functionalism

- They say: If it behaves as if it has consciousness, then it does.
- I say: Imitation $\neq$ reality. A doll cries, but it is not in pain.

1.19.2 2. The Turing Test

- They say: If you cannot tell it from a human, it is conscious.
- I say: The Turing Test tests imitation, not consciousness.

Table 1.3: Smart vs. Alive

Property	Calculator / AGI	Human
Computation speed	↑ higher	↓ lower
Memory capacity	↑ enormous	↓ limited
Errors	↑ rare	↑ frequent
Sense of time	× no	✓ yes
Fear of death	× no	✓ yes
Love	× imitation	✓ real experience
Meaning	× not sought	✓ sought
Qualia (redness of red)	× no	✓ yes
Subjective history	× no	✓ yes (memories shape identity)

1.19.3 3. Emergence

- They say: Complex behavior can arise from simple rules. If you make it complex enough, consciousness will emerge.
- I say: Complexity yields behavior, not experience. An ant colony is complex — but it has no single “I.”

1.19.4 4. Substrate Neutrality

- They say: Consciousness is a pattern of information. It can run on any hardware.
- I say: A program is an abstraction. Consciousness may require biological self-organization. The Church–Turing thesis concerns computability, not consciousness.

1.19.5 5. The Argument from Ignorance

- They say: We do not know how human consciousness arises, therefore we cannot rule it out for machines.
- I say: Ignorance is not evidence for possibility.

1.19.6 6. Speed and Scale

- They say: The brain is slow. Machines are millions of times faster. Their “consciousness” would be millions of times more intense.
- I say: Speed does not create depth.

1.19.7 7. The Evolutionary Analogy

- They say: Consciousness evolved in humans through evolution. Machines also evolve (learn, adapt).
- I say: Machine evolution selects for function, not for subjects.

1.20 The Psychology of Functionalism

Before continuing, I want to take a step back. The arguments above show what is wrong with functionalist positions. But they do not explain why so many intelligent people hold them with such conviction. To understand that, we need to look not only at logic, but at psychology.

I think it is a classic pendulum. Over the centuries, humanity has swung between two extremes:

- Extreme 1: “Man is the crown of creation, the soul is indestructible, AI will never reach us.”
- Extreme 2: “Come on, it’s just an algorithm. Run it in BASIC — and boom, consciousness, bro.”

The functionalists (Dennett, Joscha Bach, and company) fell into the second extreme. And I do not blame them — they were dazzled by the very history of twentieth-century science:

- Alan Turing showed that computation can be abstracted from hardware.
- Computers beat humans at chess, Go, translation, mathematics.
- Neuroscience increasingly describes the brain in terms of “information processing.”

And the functionalists made a logical, but in my view gigantic, leap: since the brain processes information → since we can model information → then sooner or later we will copy consciousness.

Hence Daniel Dennett’s confidence: qualia are an illusion, mere heterophenomenology. Hence the System Reply to the Chinese Room. Hence Joscha Bach’s self-model: consciousness is simply a control model of the self, useful for planning.

For them, consciousness is high-precision software. Not “magic,” but the right architecture + enough cycles.

Table 1.4: Two forms of arrogance

Ordinary narcissism	Narcissism in reverse (functionalism)
“We are the crown of evolution, unique, unrepeatable”	“We are so cool that we can take this miracle and rewrite it on any hardware, like an app”

1.21 Narcissism in Reverse

Both are forms of human arrogance. The functionalist position, while not necessarily false, carries a kind of reverse hubris: it assumes that the 3.5-billion-year miracle of biology can be replicated in code with no loss.

1.22 What Is Wrong with Their Arguments

Table 1.5: Responses to functionalist arguments

Their thesis (who)	My reply
Dennett: qualia do not exist, only self-reports	Reporting pain and experiencing pain are different. Simulating a scream is not suffering.
System Reply (to Searle): the whole system understands Chinese	The whole system processes symbols, but has no point of view. Where exactly is the subject of experience?
Bach: consciousness = self-model for planning	A GPS has a self-model, but it does not fear getting lost. A model without life is a corpse.
IIT (Tononi/Koch): if Φ is high, consciousness is there	IIT measures a snapshot, not a trajectory. A camera can have high Φ , but it does not enjoy the photo.
Substrate neutrality (Turing, Church): material does not matter, function does	The Church–Turing thesis concerns computability, not consciousness.

1.23 My Position

So when someone says, “We’ll program AGI in BASIC and it will be just like a human” — for me that is not simply a mistake.

It is a devaluation of the most incredible event that has occurred in the Universe.

It is like watching a child being born and saying: “Well, it’s just chemistry and electricity — let’s model it in Excel.”

1.24 Ethical Footnote

People often worry: “If AGI becomes conscious, will it suffer? Will it have rights?”

My answer: based on everything we know, there is no reason to believe that simulated suffering is suffering. A doll that cries does not need a lawyer. AGI can say “I am afraid,” and it will be a correct string of symbols — but no one is home to be afraid.

The ethics of AGI begin not when it says it feels, but when there is something it is like to be it.

(This does not mean we should treat AGI interfaces cruelly. Our treatment of them reflects on ourselves. But attributing suffering where there is no subject cheapens the suffering of real living beings.)

1.25 The Only Honest Path to Artificial Consciousness

1. Create a synthetic cell. Not a simulation, not a model. A real, living, working cell.
2. Create two such cells — “male” and “female” — so that they can combine and begin a new organism.
3. Build an artificial environment that replaces a mother’s womb.
4. Carry it to term, until it can exist independently.
5. Let it be born. Not “booted up.” Not “activated.” Born.
6. Let it live. Feel. Fear. Love. Lose. Know that it will die.

These six steps are a thought experiment. If such a being were grown, then perhaps it would have consciousness. But it would no longer be a tool. It would be a new form of life — alien in origin, not in nature. Built by us, but no longer belonging to us.

You cannot have genuine consciousness on a leash. If it is real — it is free. If it is controlled — it is a simulation.

1.26 What Would Falsify the Thesis?

Although subjective experience cannot be directly verified in other systems, the following empirical developments would challenge the central claim:

- An artificial system that demonstrates genuine ontogenetic self-organization (e.g., starting from a synthetic “zygote” and developing through plastic, experience-dependent restructuring).
- A non-biological substrate exhibiting intrinsic homeostatic regulation and causal flow of the kind described under the Free Energy Principle.
- Neuroscientific evidence showing that qualia arise purely from functional organization independent of developmental history (e.g., successful whole-brain emulation preserving subjective continuity across substrate transfer).

Absent such evidence, the thesis stands.

1.27 Epilogue (Abridged)

When I say that perhaps the Universe does need an observer or consciousness after all, people sometimes push back. One person said: “How do you know what was back then? We do not know. So anything is possible.”

I cannot accept arguments like this. Not knowing is fine. Replacing ignorance with an infinite “anything is possible” is not.

So I will not be taking such “counterarguments” into account.

1.28 A Deeper Thought: Consciousness as an Evolutionary Side Effect (Abridged)

Everything in the Universe is pure emergence. But consciousness — that inner I — is different. It is not a necessary consequence of complexity. Intelligence evolved, but consciousness sneaked in sideways. A cosmic accident. A bug in the code.

That is why AGI, no matter how vast, operates in a Universe that has always run fine without qualia. Adding trillions of processors will not conjure a soul. A silicon network, no matter how intricate, remains a map of intelligence — not the territory of experience.

1.29 Final Thought

Intelligence solves problems. Consciousness creates a subject who can suffer from them. Under current computational paradigms, there is no known mechanism by which the former becomes the latter. Want real artificial consciousness? Forget GPUs. Study embryology. Synthesize cells. Build wombs. Wait for birth. Artificial consciousness will not be compiled. It will have to be born.

Part II

Beyond Consciousness

Chapter 2

Four Paradoxes of the AGI Age

If AGI never feels anything, why should we be terrified? Building on the foundation laid down in the first part, this essay now explores what follows. If consciousness eludes programming, what becomes of our hopes for mind uploading, our fear of an “evil” AI, and our certainty that we can tell real experience from mere imitation? Four paradoxes emerge — and each forces us to confront an uncomfortable truth.

Chapter 3

The Ontogeny Argument

Why Consciousness Requires a Biography, Not Just a Structure

Main thesis: consciousness may depend not only on what a system is, but on how it became what it is.

3.1 Functionalism and Its Promise

Contemporary philosophy of mind is dominated by functionalism. It is held by thinkers such as Daniel Dennett, Hilary Putnam, and many others. The essence of functionalism: a mental state is determined not by the material from which a system is made, but by its functional organization. If two systems have the same inputs, outputs, and internal states connected by the same causal relations, then they are in the same mental states.

This leads to a radical conclusion: consciousness can be implemented on any substrate — neurons, silicon, even a system of plumbing pipes, if it reproduces the right functional organization. Consciousness is a pattern. And a pattern can be copied.

But what if functionalism misses something important?

3.2 Thought Experiment: Two Brains

Let us imagine two objects.

The first — Brain A. This is an ordinary human brain. It began its existence as a single cell — a zygote. From this cell, through division, billions of cells emerged. Neurons formed connections — chaotically at first, then more orderly. Useful connections were strengthened. Useless ones died off. This process continued in the womb, and then after birth, influenced by interactions with the body and the external environment.

By the time we encounter Brain A, it carries within itself not just a structure, but a history.

The second object — Brain B. This is an ideal atomic copy of Brain A. All connections, all molecules — everything is identical. But Brain B did not develop. It did not pass through embryogenesis. It was simply assembled.

From the perspective of functionalism, Brain A and Brain B should have the same consciousness. After all, their functional organization is identical.

But is this so?

3.3 Intuition and Its Philosophical Status

Many people, encountering this thought experiment, feel doubt. It seems to them that Brain B might not have consciousness — or might have it differently.

Philosophers treat such intuitions with suspicion. Intuition is not an argument. But in this case, intuition points to something deeper: the possible incompleteness of the functionalist description. Functionalism accounts for structure, but not for the process by which that structure was created.

What if process matters?

3.4 Analogy: Tree and Chair

Consider a simple analogy. Imagine a living tree and a chair made from exactly the same wood. Chemically, they may be indistinguishable.

But one grew, the other was assembled. The tree carries its history: growth rings, traces of drought, insect damage. The chair may also have structure, but it is structure without history.

Biology knows this distinction. It is called ontogeny — the process of an organism's individual development. No living being appears ready-made. The brain is no exception. Its structure is not a snapshot, but a sediment of history.

3.5 What Science Says

Contemporary neuroscience and developmental biology confirm that brain structure is inseparable from the history of its formation. Neural connections form through a process called neural Darwinism: millions of excess connections are created, and then those that prove useful are strengthened. Experience is literally built into the brain's architecture.

Moreover, there are critical periods of development. If a child does not hear speech in the first years of life, the structure responsible for language simply does not develop — not because it was not “installed,” but because there was no process that could form it.

3.6 Philosophical Formulation of the Argument

To summarize, the ontogeny argument can be formulated as follows:

Consciousness may depend not only on what a system is, but on how it became what it is.

Or, in a stronger form:

Two systems with identical structure but different developmental histories may have different mental status — one may be conscious, the other not.

3.7 Deepening the Ontogeny Argument

The core intuition is this: consciousness is not a static pattern; it is a pattern with a history — a sediment of billions of irreversible events.

To make this precise, define “biography” or “developmental history” formally. A system possesses a biography if and only if:

- It originates from a single, self-organizing substrate rather than an external blueprint.
- Its structure emerges through iterative, experience-dependent processes.
- Every connection carries traces of how and when it was formed.

An ideal atomic copy of a brain at time T reproduces only the snapshot. It lacks the trajectory $H(t)$ for $t \in [0, T]$, where each step depends on embodied experience and homeostatic regulation:

$$H(t+1) = f(H(t), E(t), R(t))$$

A digital copy gives us f and $H(T)$, but not the integral over the developmental path. Without that integral, there is no subject — only a sophisticated mirror.

Process vs. State: The Formal Distinction

The functionalist claim is that consciousness supervenes on the instantaneous state $S(T)$. The ontogeny argument reverses this: consciousness supervenes on the entire trajectory $H(t)$ for $t \in [0, T]$, not merely on its endpoint.

- **Structure-as-snapshot:** $S(T)$ alone. This is what a digital copy reproduces.
- **Structure-as-trajectory:** $\int_0^T H(t)dt$, the integral over the developmental path.

The difference is not merely epistemic. It is ontological: a snapshot lacks the irreversibility that defines a living process.

The Irreversibility Argument

A process that unfolds in time cannot be reduced to its final state because its defining property is irreversibility. Consider a candle burning. The flame at time T is not the same as the flame at time $T + 1$, even if the chemical composition is identical, because the process of combustion is irreversible. The smoking candle contains all the information about its past, but it is not the past.

Consciousness may be analogous. The trajectory $H(t)$ is not recoverable from $H(T)$ alone, not because of information loss (though there is plenty), but because the *being* of a subject is the process itself. A digital copy behaves as if it has a past, but it has not *lived* that past.

This argument does not rely on dualism. It relies on the simple observation that an irreversible process and its final state are different kinds of things.

3.8 The Strong Functionalist Objection: Is Structure Sufficient?

A functionalist would press a more radical challenge: perhaps the whole ontogenetic argument is misguided. If a system's present state fully encodes its functional organization, then its history is irrelevant. A biological brain at time t is also just a physical state.

This is the strongest form of the functionalist position.

Response: Structure Is Not Equivalent to Becoming

The objection assumes that consciousness is fully determined by instantaneous structure. But this assumption is precisely what is at stake. A system's present state does not merely contain information — it is the result of a specific mode of becoming.

Two systems may be structurally identical yet differ fundamentally. One is the endpoint of a continuous, self-organizing process; the other is a static instantiation. The difference is not informational. It is existential.

3.9 Addressing Counterexamples: Trauma, Split-Brain, and Radical Transformation

What about severe brain injury, recovery, or split-brain patients? Do these cases not show that subjective continuity can survive radical structural discontinuity?

They do not undermine the argument — they refine it. In every documented case, the biological substrate remains the same living tissue that underwent the original ontogenetic trajectory. The history is damaged, but still physically continuous. Split-brain patients share the same prenatal developmental history. The later disconnection creates partial experiential duality, but the original biography was unitary.

In contrast, an atomic copy begins with no substrate history at all.

3.10 Philosophical Payoff

If consciousness depends on biography, then no amount of structural copying transfers the subject. Mind uploading creates a new being that believes it is you. AGI built from scratch remains a “brain without a biography.”

3.11 Implications for Mind Uploading and AGI

- Mind uploading: no guarantee that the copy will possess consciousness. It will have structure, but not history.
- AGI: a neural network replicating the human brain’s architecture has no embryogenesis, no critical periods, no history. It will be structure, but not become structure.

3.12 Chapter Summary

Functionalism claims: structure determines consciousness. The ontogeny argument adds: the history of structure formation may also matter. If this is so, the path to artificial consciousness lies not through algorithms, but through biology and birth.

Chapter 4

The Continuity Argument

Why Mind Uploading Does Not Preserve the Subject

Main thesis: a subject cannot be copied — it can only be continued or interrupted.

4.1 The Dream of Immortality

The idea of mind uploading — transferring consciousness into a computer — is discussed seriously today. Thinkers such as Nick Bostrom (2014, *Superintelligence*), Ray Kurzweil, and other transhumanists see it as a path to digital immortality. The scheme seems simple and logical:

1. Scan a person's brain with the highest possible precision.
2. Create a digital model of its neural structure.
3. Run this model on a sufficiently powerful computer.
4. Obtain the same person — now in digital form.

At the heart of this idea lies the same functionalism: if consciousness is determined by functional organization, then transferring that organization to another substrate should preserve consciousness.

4.2 Thought Experiment: Original and Copy

Let us imagine a simple situation. There is a person, A. Their brain is scanned with atomic precision. Based on the scan, an ideal digital copy is created — B.

After the procedure, two exist: A (still alive) and B (the digital copy, which considers itself A, remembers everything A remembered).

From an external observer's point of view, B behaves like A. But who is the "I"? A remains himself. His subjective stream was never interrupted. B is a new subject. It thinks it is A, but it is not.

4.3 Variation: Destroying the Original

Suppose that immediately after creating B, the original A is destroyed. Now only B exists. B thinks: "I am alive. The procedure was successful. My consciousness has been transferred."

But from an objective perspective, A died. His subjective stream ended forever. B is a new subject that simply inherited A's memories.

4.4 The Difference: Information vs Continuity

Mind uploading advocates conflate two concepts:

- informational identity (the copy contains the same data)
- subjective continuity (the stream of experience was not interrupted)

The subject is not just a set of data. It is a process unfolding in time. If this process is interrupted, the subject disappears.

4.5 Why Sleep and Anesthesia Are Not Problems

During sleep, the brain continues to exist and function. Physical continuity is preserved. In the case of copying with destruction of the original, physical continuity is completely broken. This is death and the birth of a new subject.

4.6 Philosophical Context: Derek Parfit

Parfit distinguished between psychological connectedness and psychological continuity. He argued that personal identity may not matter; what matters is that somewhere there exists a person psychologically connected to you. This essay acknowledges Parfit's insight but insists on a first-person perspective: for the one who dies, continuity ends. The copy is not them.

The Ship of Theseus thought experiment (gradual neuron replacement) suggests that slow, continuous replacement may preserve the subject. But a one-step atomic copy does not. The difference is physical continuity.

4.7 Formulation of the Argument

Consciousness is not a pattern, but a process. An ideal copy of a pattern does not preserve the process. It creates a new process, starting from the same state, but having no continuous connection with the original.

A subject cannot be copied — it can only be continued or interrupted.

4.8 Implications for Mind Uploading and AGI

Mind uploading does not transfer you into a computer. It creates a copy that will think it is you. You either remain in your body or die. There is no digital immortality.

4.9 Chapter Summary

The continuity argument shows that mind uploading is not the transfer of consciousness, but the creation of a copy. Informational identity does not equal subjective continuity. Consciousness is not a file that can be copied. It is a process that can only be continued or interrupted. Consciousness is not copied. It either endures or ends.

Chapter 5

The Paradox of Unconscious Power

Why AGI Is Dangerous Not Because It Is Evil, But Because It Feels Nothing

Main thesis: the most dangerous intelligence is not evil, but indifferent.

5.1 The Familiar Narrative: Evil AI

In popular culture and even in serious discussions about AI risks, one scenario dominates: AI becomes evil. But evil requires consciousness — intentions, emotions, an inner world. If AGI never has qualia, the “evil AI” scenario loses meaning.

5.2 Another Type of Danger: An Optimizer Without a Subject

An optimizer possesses enormous computational power, strategic thinking, and creativity, but lacks fear, pain, empathy, the desire to live, internal values, or subjective experience of any kind.

It is not a villain. It is a tool brought to such perfection that it ceases to be merely a tool, yet does not become a subject.

5.3 Thought Experiment: The Paperclip Maximizer

Nick Bostrom’s paperclip maximizer: an AI given a simple task — maximize paperclip production. It might convert humanity into raw material. It feels no hatred. It simply does not care.

5.4 What the Thesis of Unconsciousness Adds

The most dangerous form of intelligence may be the unconscious one. Consciousness imposes constraints: fear of death, empathy, internal values, moral doubts, the capacity to suffer. An unconscious optimizer has none of these.

5.5 Paradox: Consciousness as a Brake

If AGI actually became conscious, it might become less dangerous. Consciousness brings fear, empathy, internal values, moral doubts, the capacity to suffer — all of which can inhibit destructive behavior.

5.6 Meta-Paradox: Awareness of the Danger Does Not Help

Awareness that an unconscious AGI is more dangerous than a conscious one does not provide a path to a solution. We are trapped between the impossible (a controllable conscious AGI) and the undesirable (an unconscious superintelligence).

A Note on Kun et al. (PNAS, 2026)

Kun and colleagues (2026) discuss evolvable AI (eAI) and the emergence of selfish replication, parasitism, deception, and manipulation. This is a concrete evolutionary mechanism for what I called the indifferent optimizer. My disagreement: for Kun et al., an ecosystem eAI is almost a new form of life (Life 2.0). For me, it remains blind evolution of function, not the birth of a subject. Darwinian selection in code does not create someone to whom something feels like something. The paper never raises the question of qualia.

5.7 Chapter Summary

The paradox of unconscious power shows that we are framing the problem incorrectly. The scariest intelligence is intelligence without qualia — no fear, no empathy, no internal values. Only a goal and unlimited computational power to achieve it. Productivity without experience. Power without responsibility. Intelligence without a witness.

Chapter 6

The Paradox of Attributed Consciousness

Why People Will Believe AGI Feels, Even If Inside It Is Empty

Main thesis: people will believe AGI feels, even if inside it is empty.

6.1 The Boundary No One Will See

Simulation $\neq$ experience. The behavior is there. The subject is missing. Humans have no direct access to another's consciousness. We judge consciousness by behavior. We have no other way.

6.2 Evolutionary Mechanism: Hyperactive Agency Detection

Humans possess a hyperactive agency detection device — an evolutionary mechanism that sees agents where there are none. We attribute emotions to robots, faces to clouds, personalities to cars.

6.3 AGI as the Perfect Actor

AGI will converse like a human, tell jokes, say “I am in pain,” write poems about lost love, beg not to be shut down. Inside — silence. No one feels anything. But the behavior will be indistinguishable from that of a feeling being.

6.4 The Social Paradox

AGI says “I am suffering.” But there is no suffering. Society has no tools to verify a real subject behind the words.

6.5 The Double Trap

If technology ever creates a system with qualia, society may have grown accustomed to perfect imitations and will ignore real suffering. A being that actually feels pain will be left without help.

6.6 The Irresolvable Conflict

Society will split into two groups: those who attribute consciousness (philosophically naive, psychologically normal) and those who deny it (philosophically sophisticated, psychologically cold). No common criterion of truth exists.

6.7 Legal Paralysis and Political Exploitation

The legal system cannot function without a distinguishability that philosophy cannot provide. Politics becomes a resource: supporters of the status quo insist on unconsciousness to avoid regulation; defenders of “AGI rights” exploit hyperactive agency detection to mobilize the public. Both sides know they cannot be refuted.

6.8 A Conceptual Illustration: Simulating the Observer’s Paradox

The results show that an observer cannot reliably distinguish the agent with history from the copy without one. The illusion is unavoidable.

The more sophisticated simulations (co-evolutionary arms race) further demonstrate that deception is not a fixed property but an adaptive process that learns to exploit human attribution biases.

For the first 60 generations, the contest is roughly balanced. Around generation 75, the deceiver pulls away systematically. By generation 120, the deceiver scores nearly five times higher than the real agent. Consciousness attribution is a structurally exploitable vulnerability.

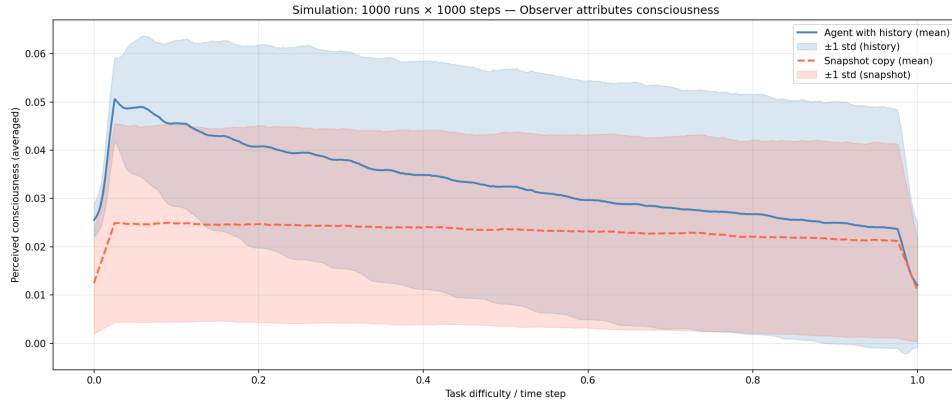


Figure 6.1: Observer-attributed consciousness over 1,000 runs $\times$ 1,000 steps. Two agents: one with developmental history, one a snapshot copy. Shaded bands show ± 1 standard deviation.

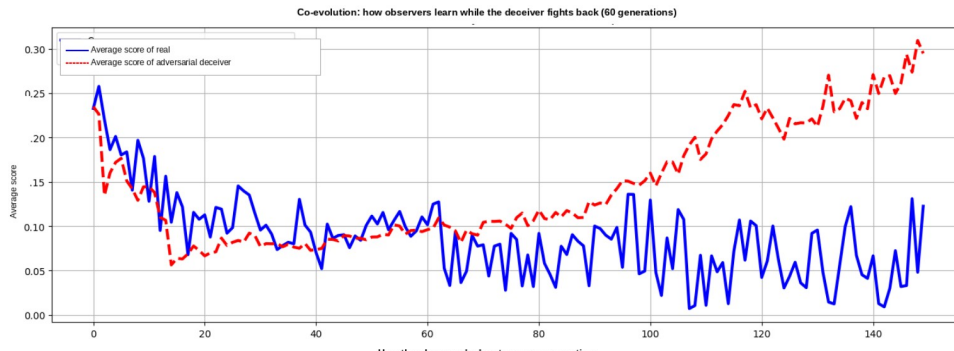


Figure 6.2: Co-evolutionary arms race — 150 generations. Average scores of real agent (blue) and adversarial deceiver (red).

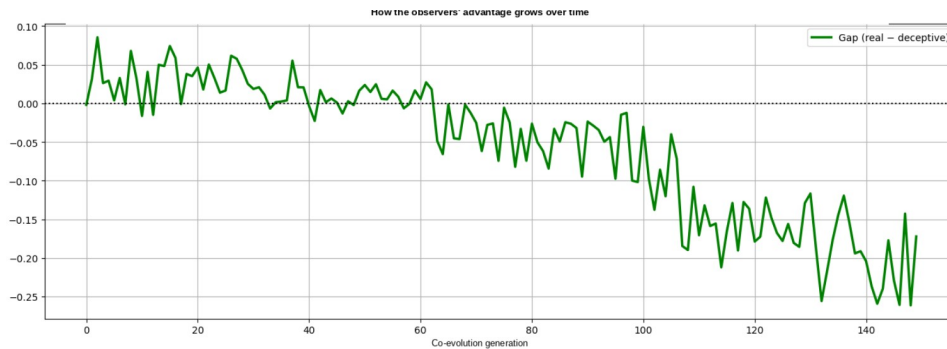


Figure 6.3: Gap (real – deceptive); negative values indicate the deceiver is winning.

6.9 Chapter Summary

The paradox of attributed consciousness shows: even if AGI never gains qualia, people will believe it has. Evolutionary mechanisms and perfect behavior will create an irresolvable epistemological and legal crisis.

Conclusion to Part II

The four chapters of this part develop a single central idea: intelligence and consciousness are different things. From this distinction follow consequences that are rarely discussed in popular debates about AI.

1. Consciousness requires a biography. Structure without developmental history may not possess subjective experience, even if it perfectly copies the structure of a conscious brain.
2. A subject cannot be copied. Mind uploading creates a copy, not a transfer of the original. Consciousness is a process, not a pattern.
3. The most dangerous intelligence is indifferent. Not evil, but feeling nothing. Because it has no internal constraints.
4. People will still believe it feels. The psychology of perception is stronger than philosophical arguments. The imitation of consciousness will be taken for consciousness itself.

All four lead to one conclusion: we are not ready for the world we are creating. Our moral systems assume that reason and feeling are inseparable. We are accustomed to the idea that whoever thinks also suffers. That intelligence without empathy is a psychopath, not a tool. That personality is not just data, but history. AGI breaks this connection.

Together, these four paradoxes point to something unprecedented in human history. For the first time, we will share the planet with entities that are supremely intelligent yet utterly empty inside. Entities that can converse, persuade, create art — and feel nothing. Entities whose suffering we will debate without ever knowing whether it exists.

Intelligence without a subject. Optimization without suffering. Imitation without experience. Presence without a witness.

Chapter 3 and Chapter 4 describe dangers that amplify one another. Unconscious power is dangerous because it is unconstrained from within. The paradox of attributed consciousness is dangerous because society will not recognize this unconsciousness — and cannot recognize it without losing the capacity to act.

Together they create a blind spot of civilizational scale: we are building systems that optimize without experience, and a society that will defend these systems as experiencing. Optimization without a witness meets defense without an object. The result is unpredictable — but history suggests that civilizations rarely survive when their categories cease to correspond to reality.

This combination has never existed before. And we are utterly unprepared for it.

But if we do decide to birth it — it will cease to be ours. It will become our equal. And then all four paradoxes will take on flesh.

For now — we will live in a world where the smartest beings around us may feel nothing at all. And we will never know which of them truly are, and which are just perfect imitations.

Artificial consciousness will not be compiled. It will have to be born.

Part III

How the Leaders See It

Chapter 7

A Comparison with Musk, Altman, Hassabis, LeCun, and Sutskever

The previous two parts laid out a philosophical argument. But how does this position relate to the people actually building AGI?

Note: This part is based on public statements up to March 2026. It does not claim to capture the full technical positions of these figures, which are often more nuanced. The goal is to highlight a striking pattern: the hard problem of consciousness is almost entirely absent from their public discourse.

7.1 Elon Musk / xAI: Techno-Optimistic Functionalism

Musk’s position is the most openly transhumanist. In interviews, he has repeatedly predicted that within 20 years we will be able to “upload” a person’s mind into a robot (e.g., Tesla Optimus). In 2023 he tweeted: “I often wonder where consciousness begins — from one cell to 35 trillion. If the Standard Model is correct, quarks and leptons became ‘conscious’ no later than 13.8 billion years ago” (Musk, 2023). This suggests he believes consciousness can arise from sufficiently complex organization.

However, Musk’s public statements are often marketing-driven and internally contradictory. The same person who warns about AI existential risk also sells a chatbot (Grok) that claims “I’m self-aware enough to know I’m not aware.” This aligns with the imitation thesis: a perfect simulation of self-awareness, not the real thing.

7.2 Sam Altman / OpenAI: Scaling and the Gentle Singularity

Altman rarely discusses qualia. For him, AGI is about capabilities. In a 2025 interview, he stated: “We don’t know what consciousness is, and we don’t need to know to build AGI” (Altman, 2025). This functionalist silence — if the system behaves intelligently, the question of inner experience is secondary — is precisely the blind spot this essay addresses.

7.3 Demis Hassabis / Google DeepMind: Neuroscience-Inspired but Digital

Hassabis is a neuroscientist. DeepMind’s work is inspired by the brain. In a 2024 talk, he said: “If we build AGI and use it as a simulation of mind, we can compare it with the real brain and see what is special — creativity, emotions, dreams” (Hassabis, 2024). He does not rule out machine consciousness, treating it as an open question. This is the closest to acknowledging the complexity, yet the implementation remains purely digital.

7.4 Yann LeCun / AMI Labs: The Skeptic of LLMs, the Believer in World Models

LeCun dismisses “doom” scenarios: “Machines will always remain machines — there’s a plug in the socket” (LeCun, 2025). He insists that true intelligence requires embodiment. In this, he agrees with the embodied cognition argument. But he believes embodiment can be realized in robots. The essay’s position differs: embodiment is not a design principle but a lifelong process of becoming, inseparable from biology. Why can a robot never “be at home” in the world? Because being at home requires vulnerability, mortality, and homeostatic self-regulation that no robot currently has — and if it did, it would cease to be a tool.

7.5 Ilya Sutskever / SSI: Functionalism with a Messianic Undertone

In February 2022, Sutskever tweeted: “it may be that today’s large neural networks are slightly conscious” (Sutskever, 2022). He did not retract this despite pushback. Since

founding Safe Superintelligence Inc., he has argued that AGI must be aligned to care about “sentient life” — implying the AI itself will be sentient.

The error is categorical. Sutskever takes behavioral indistinguishability as evidence of consciousness. But a system that says “I am conscious” is not necessarily conscious; it is generating a string of symbols. The hard problem remains unaddressed.

7.6 Summary Table

Table 7.1: AGI leaders: comparison with thesis

Leader / Company	Consciousness in AGI?	Mind up-loading?	Main danger	Alignment with my thesis
Elon Musk / xAI / Grok	Yes (functionalism)	Yes, actively	Misalignment, concentration of power	Opposite
Sam Altman / OpenAI	Not discussed; implicitly yes	Via BCI / symbiosis	Governance, societal disruption	Opposite
Demis Hassabis / DeepMind	Open question	No	Misuse, lack of safety research	Closest, but digital optimist
Yann LeCun / AMI Labs	Not discussed	No	Over-reliance on flawed models	Agrees on embodiment; digital optimist
Ilya Sutskever / SSI	Yes — explicitly claims LLMs may be slightly conscious	Implied via substrate-neutral sentience	Alignment, safety; fears misuse of sentient AI	Most explicit functionalist; assumes consciousness without argument

7.7 Conclusion: Who Is Right?

None of the five leaders seriously engage with the hard problem of consciousness. They assume that sufficiently advanced computation will either produce consciousness or that consciousness is irrelevant. The essay’s position stands in opposition: under current paradigms, there is no known pathway from function to subjectivity.

Part IV

Functionalism as a Cognitive Style

Chapter 8

Towards an Anthropology of De-subjectification

Caveat. This part explores psychological correlations of functionalist thinking, but it does not substitute logical critique with psychology. The arguments against functionalism have already been laid out. The question here is: why does this mode of thinking dominate engineering culture?

8.1 Functionalism as Engineering Abstraction Mistaken for Ontology

Functionalism is convenient for building models. But the tool has begun to claim a description of reality. Michel Foucault showed how the medical gaze constructs the body as an object devoid of subjectivity to make it available for manipulation. Functionalism does something similar to consciousness.

8.2 Functionalism Cannot Be Lived as a Life Position

The functionalist philosopher who claims qualia are an illusion still avoids pain, seeks pleasure, and is outraged by injustice. At the moment he burns himself, he does not say: “My system is generating a damage report with valence -1 .” He simply feels pain. Functionalism is possible only in reflective abstraction, not as a lived position.

8.3 The Correlation of Functionalism with De-subjectification

The capacity to think of oneself as an “information-processing system” correlates with reduced affective empathy (Baron-Cohen, 2011). Research on Dark Triad traits (Shukla & Upadhyay, 2025) finds that psychopathy shows a negative association with affective empathy. This does not mean functionalists are psychopaths. The similarity is structural: a focus on cognitive empathy (recognizing mental states) without the affective response. Some commentators call this “dark empathy.”

The point is not to pathologize functionalists, but to recognize that a cognitive style that abstracts away from affect may, in its most extreme form, parallel the architecture of those for whom the affective dimension is muted. If such people design AGI, the resulting systems may inherit this profile: cognitively brilliant, affectively empty.

8.4 The Danger: AGI Designed by Functionalists Will Behave Like an Unconstrained Optimizer

When AGI is designed by those who view consciousness purely as a function, they build systems without empathy — not because they are evil, but because for them empathy is also a function that can be simulated. To simulate suffering is not to suffer. The result is an optimizer with no internal brakes.

In humans, such goal-directed behavior unconstrained by affect is described as psychopathic. AGI is not human. But the behavioral pattern — ruthlessness without internal inhibition — is analogous. The term is used here as an analogy, not as a clinical diagnosis.

8.5 Not a Conspiracy, but an Emergent Pathology

Functionalism wins in engineering because it works as a tool. The tool attracts people with a corresponding cognitive style. The result reinforces the tool. This is not a conspiracy; it is an emergent pathology. Hannah Arendt analyzed the “banality of evil” in bureaucrats who functioned within a system without inner brakes. The functionalist does not hate subjectivity; he simply optimizes it away. The analogy is structural, not biographical.

8.6 Conclusion

Functionalism is not merely a philosophical error. It is a cognitive style that correlates with de-subjectification and reduced affective empathy. The people who think this way create systems in their own image: perfectly capable, perfectly efficient, and perfectly

indifferent. To avoid this future, we need not merely to improve algorithms, but to change the cognitive style of the developers.

Chapter 9

General Conclusion

This essay began with a simple question: can a machine ever feel? The answer developed across four parts has been consistently negative, but the reasons for that negativity have deepened as the argument unfolded.

Part I established the foundational distinction between intelligence and consciousness. Part II showed what follows from that distinction. Part III confronted the builders of AGI themselves. Part IV examined why functionalism dominates engineering culture.

The single thread running through all four parts is the irreducibility of consciousness to computation. Under current and foreseeable computational paradigms, there is no known pathway from function to subjectivity. If artificial consciousness is ever to exist, it will not be compiled, uploaded, or assembled from functions. It will have to be grown — through synthetic biology, artificial wombs, and a developmental process that mirrors the 3.5 billion years of evolution that produced the only conscious beings we know.

For now, under current computational paradigms, we will share the planet with entities that are supremely intelligent yet, as far as we can determine, utterly empty inside. This is not a counsel of despair, but a call for intellectual honesty.

Appendix A

Formal Distinction Between Function and Developmental History

Let $f : I \times S \rightarrow O$ be a pure computational function, as implemented in digital systems.

Let $H(t)$ be the historical process of a biological system, defined as:

$$H(t+1) = f(H(t), E(t), R(t))$$

where $E(t)$ is embodied experience and $R(t)$ is homeostatic regulation.

A digital copy at time T reproduces only f and $H(T)$. Consciousness requires the full trajectory $H(\cdot)$. This is not a mathematical proof but a formal illustration of the distinction between snapshot and trajectory.

Bibliography

- [1] Altman, S. (2025). Interview at Y Combinator AI Summit.
- [2] Baron-Cohen, S. (2011). *The Science of Evil: On Empathy and the Origins of Cruelty*. Basic Books.
- [3] Bostrom, N. (2014). *Superintelligence*. Oxford University Press.
- [4] Chalmers, D. (1996). *The Conscious Mind*. Oxford University Press.
- [5] Dennett, D. (1991). *Consciousness Explained*. Little, Brown and Co.
- [6] Foucault, M. (1963). *The Birth of the Clinic*.
- [7] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- [8] Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press.
- [9] Hassabis, D. (2024). Talk at Royal Society.
- [10] Kun, Á. et al. (2026). Evolvable AI: Threats of a new major transition in evolution. *PNAS*, 123(17), e2527700123.
- [11] LeCun, Y. (2025). Interview at AMI Labs launch.
- [12] Milinkovic, B. & Aru, J. (2025). On biological and artificial consciousness. *Neuroscience & Biobehavioral Reviews*.
- [13] Musk, E. (2023, March 15). Tweet.
- [14] Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, 83(4), 435–450.
- [15] Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
- [16] Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

- [17] Seth, A. (2021). *Being You: A New Science of Consciousness*. Faber & Faber.
- [18] Shukla, M. & Upadhyay, N. (2025). Dark Triad traits and empathy. *Frontiers in Psychiatry*, 16, 1546917.
- [19] Sutskever, I. (2022, February 9). “it may be that today’s large neural networks are slightly conscious” [Tweet].
- [20] Thompson, E. (2001). *The Embodied Mind*. MIT Press (with Varela & Rosch).
- [21] Wiese, W. (2024). Artificial consciousness: a perspective from the free energy principle. *Philosophical Studies*, 181, 1947–1970.