§ THE DIALECTIC REVIEW

Volume III · Spring 2025

ISSUE 05

Minds and Machines

Spring 2025 · 7 pieces

Seven essays on cognition, simulation, and what, if anything, separates a mind from a very good model of one. This issue predates The Counterargument, introduced in Issue 06.

ISSN 3049-7154

thedialecticreview.org

CONTENTS

PHILOSOPHY OF MIND 3

What a Thermostat Doesn't Know

Henrik Solberg — Fjellheim Videregående Skole, Bergen

PHILOSOPHY OF MIND 6

The Chinese Room, Twenty Years Later

Layla Haddad — Greenfield International School, Amman

AESTHETICS 9

Do Algorithms Deserve Authorship?

Beatriz Alcantara — Colégio Marista, Curitiba

METAPHYSICS 12

The Ship of Theseus Runs on Software

Wen Jing Ho — Chinese International School, Hong Kong

ETHICS 15

Why Simulated Pain Might Still Hurt Someone

Samuel Okoro — Greenwood Comprehensive, Port Harcourt

PHILOSOPHY OF SCIENCE 18

What Turing Actually Bet On

Isabel Torres — Instituto Nacional, Santiago

METAPHYSICS 21

Minds Without Selves

Petra Lindgren — Sankt Olof School, Stockholm

PHILOSOPHY OF MIND

What a Thermostat Doesn't Know

A thermostat represents the temperature and acts on that representation. This essay asks what more, exactly, a mind is doing that a thermostat is not.

Henrik Solberg Fjellheim Videregående Skole, Bergen

A thermostat takes in information about room temperature, compares it to a set point, and triggers heating or cooling to close the gap. Described this way, it is hard to avoid a certain vocabulary: the thermostat "wants" the room to be a certain temperature, "believes" the room is currently some other temperature, and "acts" to bring the two into alignment. Philosophers have used exactly this example, going back to John McCarthy and later popularized by John Haugeland and others, to ask an uncomfortable question: if wanting and believing can be fully cashed out in terms of information states and the behavior they produce, what stops us from saying, quite literally, that thermostats have beliefs and desires — extremely simple ones, but beliefs and desires nonetheless? This essay takes the question seriously rather than dismissing it, and argues that the right answer is not "thermostats believe nothing" but "belief comes in a genuine hierarchy, and the thermostat sits at the bottom rung of something continuous, not a different thing entirely."

The temptation to dismiss the thermostat case usually rests on an appeal to complexity: surely real belief requires a much richer information-processing system, capable of holding many

representations at once, revising them in light of new evidence, and combining them in flexible ways the thermostat cannot. This is true, but notice that it is a difference of degree — more representations, more flexibility, more integration — not a difference that identifies some special ingredient present in minds and absent in thermostats. If richness of information processing is what separates real belief from the thermostat's pseudo-belief, then belief is not a binary property some systems have and others lack; it is a matter of how much of a certain kind of structure a system possesses, and the thermostat has a nonzero, if minimal, amount of it.

A more promising place to look is not at complexity but at a specific structural feature philosophers call functional role: a genuine belief is not just a state correlated with the world, but a state that interacts with a wide variety of other states — other beliefs, desires, plans — in ways sensitive to their content. The thermostat's temperature-representation does exactly one thing: it feeds into exactly one comparison, producing exactly one kind of output. It cannot combine with a representation of, say, the cost of electricity to produce a more nuanced action; it cannot be held in doubt, weighed against conflicting information, or revised in light of a reason. A genuine belief system is one where representations are, in the jargon, *poised* to interact with indefinitely many other representations in ways that depend on their specific content — the belief that it is raining can combine with the desire to stay dry and the belief that the umbrella is by the door to produce a specific action, and can just as easily combine with different desires and beliefs to produce entirely different actions. This flexible, combinatorial availability is what the thermostat's state structurally lacks, not as a matter of it having too little of it, but as a matter of its architecture not supporting this kind of interaction at all.

This suggests a more precise answer to the opening question: the thermostat does not have a primitive belief sitting at the bottom of a hierarchy that includes human belief at the top. It has something that resembles belief along one dimension — carrying information that guides action — while categorically lacking the dimension that does most of the philosophical work we want the concept of belief to do, namely, combinatorial integration with other states. Richness is not irrelevant, but the relevant kind of richness is not "how much information" but "how combinable is this state with everything else the system represents." A chess engine that considers millions of positions per second has vastly more informational richness than the thermostat, but if its board-representations cannot combine with, say, a representation of its opponent's time pressure to produce a strategic bluff outside its programmed evaluation function, it may be closer to the thermostat on the dimension that matters than its sheer computational richness would suggest.

This leaves open, deliberately, whether any existing artificial system has crossed into full combinatorial integration, and by how much — that is an empirical and architectural question this essay is not equipped to settle for any specific system. What it offers instead is a better question to ask than "is this system complex enough to have beliefs," which invites exactly the sliding-scale confusion the thermostat case exploits. The better question is whether a system's internal states are poised for the specific kind of promiscuous, content-sensitive combination that separates a state that merely correlates with the world from a state that can be reasoned with — and that question, unlike a raw complexity count, at least tells us what we would need to look for.

PHILOSOPHY OF MIND

The Chinese Room, Twenty Years Later

Searle's famous thought experiment is usually treated as settled by its critics. This essay argues the strongest reply to Searle was available all along, and it is not the one most commonly given.

Layla Haddad Greenfield International School, Amman

John Searle's Chinese Room asks us to imagine a person who does not understand Chinese, locked in a room, manipulating Chinese symbols according to a rulebook so skillfully that fluent Chinese speakers outside the room believe they are corresponding with someone who understands the language. Searle's point was that the man in the room is doing exactly what a computer program does — syntactic symbol manipulation — and manifestly does not understand Chinese, however convincing the output; therefore, Searle concluded, no amount of syntactic symbol manipulation, however sophisticated, constitutes genuine semantic understanding, and this applies to any digital computer running any program, no matter how it behaves.

The most common reply, the "systems reply," concedes that the man does not understand Chinese but insists the whole system — man, rulebook, room, symbols together — does. Searle's response to this reply is genuinely forceful: have the man memorize the entire rulebook and do the symbol manipulation in his head, dispensing with the room and the papers. Now there is no separate "system" distinct from the man; he simply is the whole system, manipulating symbols in his head exactly as before, and he still, by stipulation, does

not understand Chinese. The systems reply seems to evaporate under this modification, and this is generally treated as a decisive point in Searle's favor.

But consider what this modified case is actually asking us to imagine, taken at face value: a single human brain running two entirely separate cognitive processes side by side — the man's ordinary English-speaking understanding, and a vast, memorized rule-following process sophisticated enough to conduct fluent, contextually appropriate conversation in Chinese — with the second process producing no trace of understanding accessible to the first. This is not a small ask. The rulebook, to produce fluent, contextually sensitive Chinese conversation across arbitrary topics, would need to encode something functionally equivalent to genuine linguistic and world knowledge, updated and cross-referenced in real time, at a level of sophistication that plausibly requires the exact same kind of flexible, combinatorial representation-manipulation we call understanding when we see it running in the brain's language centers. Searle's thought experiment asks us to grant that this process runs, in full, inside a single mind, without producing any understanding whatsoever — and then treats our intuition that "the man doesn't understand Chinese" as decisive evidence about the process, rather than asking whether that intuition is actually tracking the process's sophistication or merely tracking the man's own first-person, introspective report.

This is, I think, the strongest reply available, and it is different from the systems reply: it is not a claim that some larger system understands where the individual does not; it is a challenge to whether the thought experiment's central intuition pump — "clearly the man does not understand" — is actually answerable by introspection at all, once the process being asked about is stipulated to be as sophisticated as fluent, unbounded Chinese conversation

requires. A split-brain patient's left hemisphere frequently cannot report on, or introspect, information the right hemisphere is demonstrably processing and acting on; the absence of a unified introspective report is not, in the split-brain literature, taken as proof that no understanding is occurring in the hemisphere that cannot report it. Searle's argument requires us to trust the man's introspective report ("I don't feel like I understand Chinese") as authoritative over a separate cognitive subsystem's actual processing, when we already have empirical reason, from split-brain cases alone, to think introspective access of this kind is not reliable evidence about what a separate subsystem within the same skull is or is not doing.

None of this proves that the memorized rulebook process understands Chinese — that remains a live and difficult question. What it shows is that Searle's modified thought experiment does not settle the question the way it is usually taken to, because it smuggles in an assumption — that a single subject's introspective report is authoritative about all cognitive processing occurring in that subject's brain — that independent evidence about split cognition gives us specific reason to doubt. The Chinese Room is not wrong to worry about the gap between syntax and semantics. It is wrong to think the man's shrug settles which side of that gap his own, stipulated, sufficiently sophisticated symbol-processing actually falls on.

AESTHETICS

Do Algorithms Deserve Authorship?

Copyright law is being forced to decide whether a generative system can be an author. This essay argues the legal question and the aesthetic question are being wrongly treated as the same question.

Beatriz Alcantara Colégio Marista, Curitiba

Courts in several jurisdictions have recently had to decide whether a work produced by a generative algorithm, with minimal human involvement beyond a prompt, can be copyrighted, and to whom authorship should be attributed if so. The legal reasoning in these cases has, understandably, focused on questions copyright law already has vocabulary for: originality, the necessity of human creative control, the definition of "author" in existing statute. This essay argues that a separate question is being quietly resolved as a side effect of these legal debates — an aesthetic question about what authorship consists in — and that conflating the legal and aesthetic questions produces bad answers to both.

The legal question is, at bottom, a policy question: what allocation of property rights over generated works best serves the purposes copyright law exists to serve, chiefly incentivizing the creation and distribution of creative work. This can be answered without ever settling whether the algorithm itself did anything we should call authorship in an aesthetic sense; a court could reasonably decide that copyright should vest in the prompt-writer purely as a matter of who needs the incentive, or that no one owns the output, purely as a matter of what best serves public access, without this decision

implying anything about whether the algorithm exercised aesthetic judgment. Yet public debate about these cases constantly slides from "who legally owns this" to "was this really created, in the meaningful sense" – as though the legal verdict answers the aesthetic one.

The aesthetic question is harder and more interesting, and it is not settled by noting that the algorithm followed a training process rather than a lived human life, since this proves too much: a photographer who sets a camera on a tripod, chooses an exposure, and presses the shutter is also, in a sense, letting a mechanical process – light hitting a sensor according to fixed physical law – do the actual image-formation, and no one doubts photographs can involve real authorship. What we credit the photographer with is not the physics of light capture but the selection: of scene, framing, moment, exposure, among a vast space of alternatives, guided by an aesthetic sense of what would be worth capturing. The question for generative systems, stated properly, is whether an analogous selection is occurring, and where.

Here the honest answer is: it depends enormously on the workflow, and lumping all algorithmically-assisted work into one category obscures more than it reveals. A user who submits a single, generic prompt and accepts the first output is doing something aesthetically thin – closer to placing a lottery order than exercising the photographer's selective judgment, because almost no discriminating choice among alternatives has occurred. A user who iterates extensively, curating dozens of variations, refining a prompt with the same discriminating attention a photographer brings to choosing among contact sheets, is doing something aesthetically much closer to the photographer's selection, even though the underlying generative mechanism, considered alone, is identical in both cases. Authorship, on this view, was never a fact purely about the causal

mechanism producing the pixels or the words; it is a fact about where, in the whole process, discriminating aesthetic selection actually occurred, and by whom.

This means the aesthetic question cannot be answered categorically for "AI-generated work" as a class, any more than "mechanically produced work" could be answered categorically for photography as a class — some photographs are authored in the full sense and some snapshots are not, and the same distinction cuts across generative outputs depending on the actual selective work performed by the human involved. Courts, forced to draw a categorical legal line for practical reasons, will inevitably flatten this distinction in one direction or another, and that flattening is a reasonable cost of having workable law. But treating the resulting legal line as though it had settled the aesthetic question — as though a court's ruling on property rights tells us whether discriminating aesthetic judgment was exercised in a given case — mistakes a necessary simplification for an actual answer, and the actual answer, aesthetically, was never going to be the same for every case the legal ruling has to treat alike.

METAPHYSICS

The Ship of Theseus Runs on Software

Software updates replace a program's components continuously, much as the Ship of Theseus's planks were replaced one by one. This essay asks whether the classical puzzle actually gets harder, not easier, once the parts being replaced are lines of code.

Wen Jing Ho Chinese International School, Hong Kong

The Ship of Theseus asks whether an object that has had every one of its original parts replaced, gradually, over time, remains the same object. Philosophers have proposed various solutions — identity tracks spatiotemporal continuity, identity tracks a persisting form or function regardless of matter, identity is simply indeterminate past a certain point of replacement — and the puzzle is often treated as a curiosity with limited practical stakes, since actual ships rarely have every plank replaced in a philosophically clean, one-at-a-time sequence. Software, this essay argues, makes the puzzle unavoidable and practically consequential, because continuous, piecemeal replacement is not a hypothetical edge case for software — it is the normal condition of every actively maintained program, and we need an answer, not just an intuition, for questions that actually come up: is the banking app on your phone, after eight years of updates that have replaced every original line of code, the same program you first downloaded, for purposes of licensing, liability, and trust?

The spatiotemporal continuity solution, which works reasonably well for physical ships (there is a continuous physical object at every moment, even as its matter turns over), has no clean application to software, because a program is not identical to the physical hardware storing it — the "same" program can be moved to new hardware, backed up, restored, and run on an entirely different machine without anyone thinking a new, numerically distinct program has been created. If physical continuity is not what makes updated software "the same program," something else has to be doing the work, and the natural candidate is functional or structural continuity: the persistence of an architecture, a purpose, a continuous chain of version history, even as every individual function is eventually rewritten.

But functional continuity turns out to be a much weaker anchor for software than it first appears, because software identity claims routinely survive not just component replacement but functional replacement too: a messaging app that replaces its entire encryption scheme, its user interface, and its core data model across major version updates is still called, by its maker and its users, "the same app," even though almost nothing about how it functions internally, or even what it does for the user, need have survived from version one. What has survived is neither physical continuity nor functional continuity in any deep sense, but something more like a maintained institutional commitment: the continuous chain of a development team (or its successors) treating each version as an update to the prior one, embedding that claim in a persistent name, a version history, and a user base's ongoing relationship with the product, rather than launching a competitor and admitting the old thing ended.

This suggests that software identity is not being settled by any of the metaphysical criteria the Ship of Theseus literature usually proposes, but by something closer to a social and institutional fact: a program is

the same program across updates because, and only because, the relevant maintaining institution and user community continue to treat it that way, through naming, versioning, and continuity of use, not because some persisting essence — physical, functional, or structural — has been identified and verified to survive. If this is right, it has an uncomfortable implication for the classical puzzle: the physical Ship of Theseus may have looked hard because we were searching for an objective criterion that a socially-maintained identity claim was providing all along, invisibly, through the ship's name and its owners' continuous treatment of it as one vessel, and software's more extreme case of gradual replacement is not a harder version of the same puzzle so much as a case where the invisible, doing-all-the-work social criterion has become impossible to overlook, because there is nothing else left, in a fully replaced codebase, that even pretends to be doing the identity-preserving work instead.

This does not resolve the metaphysical question so much as relocate it: if social and institutional continuity is really what grounds identity claims in these cases, the interesting further question — for software, for ships, and arguably for organisms and persons too — is what makes an institutional or social continuity claim itself well-founded rather than arbitrary, since a company could, after all, simply declare an unrelated new product "version nine" of an old one, and something in us resists this as a matter of manipulation rather than genuine continuation. That resistance suggests the social criterion is not infinitely elastic — it has its own conditions of adequacy — and finding them, rather than finding a persisting plank or a persisting function, may be where the real content of the Ship of Theseus puzzle was hiding all along.

ETHICS

Why Simulated Pain Might Still Hurt Someone

Even a system with no capacity for suffering can still be the occasion of a real moral wrong, if the wrong lands on the humans who watch it.

Samuel Okoro Greenwood Comprehensive, Port Harcourt

Debates about whether artificial systems can suffer usually assume that if the answer is no, then nothing morally troubling is happening when such a system is depicted, in convincing detail, being tortured, begged with, or made to suffer for entertainment or research. This essay argues that this assumption moves too fast, because it treats the artificial system as the only possible locus of harm in the interaction, when a second, entirely uncontroversial locus of harm — the human participants and observers — is available even if the system itself feels nothing at all.

The relevant precedent is not other minds but violent fiction, and the disanalogy usually offered — "we already accept realistic depictions of suffering in film and literature without moral panic" — actually points toward the concern rather than away from it, once we notice what does worry us about certain kinds of realistic depiction even in fiction. Extended, interactive cruelty, where a human participant is not merely watching a scripted scene but personally, repeatedly choosing to inflict convincing suffering on a responsive, humanlike system that reacts in real time to that specific participant's specific choices, is structurally different from watching an actor perform a scene whose outcome does not depend on the viewer's ongoing

decisions. The psychological literature on moral disengagement and habituation suggests, plausibly though not yet definitively, that repeated practice at treating a convincingly suffering, responsive other as a legitimate target for cruelty can degrade a person's dispositions toward real suffering others, independent of whether the practice target actually suffers — in the way that repeated practice at any skill, prosocial or antisocial, tends to strengthen the corresponding disposition through simple behavioral reinforcement.

If this is right, we do not need to resolve the hard question of machine consciousness to identify a real ethical stake in how convincingly-suffering artificial systems are treated, because the stake can be located entirely in the human participant's own moral formation, which is a much less contested category than machine sentience. A parent who is unbothered by their child killing videogame enemies but uneasy about a game that rewards prolonged, personalized cruelty toward a begging, weeping, textually responsive character is not confused about where moral status resides; they are tracking a real difference in what kind of behavioral rehearsal the interaction constitutes, regardless of what, if anything, is happening on the other side of the screen.

This argument has real limits that need stating precisely, because it would be too strong, and clearly false, to say that any simulated suffering degrades the people who witness it — catharsis theories of fiction, going back to Aristotle, have long argued the opposite is often true, that engaging with depicted suffering in a controlled, narratively meaningful context can build empathy rather than erode it. The distinction that matters is probably something like this: fiction that depicts suffering as part of a story the audience does not control, or interactive fiction where the audience's choices are constrained toward morally legible outcomes, is doing something different from an unconstrained, repeatable, personalized infliction

of convincing suffering with no narrative frame giving the cruelty meaning or consequence. The former engages moral imagination; the latter risks becoming simple rehearsal.

None of this settles where exactly the line falls, and reasonable people building and using these systems will disagree about specific cases. What it does establish is that "the system cannot suffer" is not, by itself, a sufficient answer to whether a given interactive design is ethically fine, because it leaves out the party whose moral status was never in doubt: the human on the other end of the interaction, whose dispositions are being shaped by what they are permitted, encouraged, or rewarded for doing to something that behaves, in every observable respect, like it is being hurt.

PHILOSOPHY OF SCIENCE

What Turing Actually Bet On

The Turing Test is usually treated as a proposed definition of machine intelligence. Rereading Turing's original paper suggests he was making a narrower and more interesting claim.

Isabel Torres Instituto Nacional, Santiago

Alan Turing's 1950 paper "Computing Machinery and Intelligence" is almost always summarized as proposing a test for machine intelligence: if a machine can converse indistinguishably from a human, via text, it should be considered intelligent. This summary has generated seventy-five years of objections — the test rewards deception, it is anthropocentric, it conflates behavior with underlying capacity — nearly all of which assume Turing was making a metaphysical claim about what intelligence consists in. Rereading the paper's actual structure, this essay argues Turing was doing something narrower and, properly understood, harder to object to: replacing an unanswerable question with an answerable one, not defining the unanswerable question away.

Turing opens the paper by explicitly setting aside the question "can machines think" as "too meaningless to deserve discussion," given how poorly defined both "machine" and "think" are in ordinary usage. This is not a minor throat-clearing remark; it is the paper's actual thesis-setting move, and it means the imitation game was never proposed as an analysis of what "thinking" means, in the way a philosopher might propose that knowledge is justified true belief. Turing is explicit that he is substituting a different, more precisely

specifiable question — can a machine's conversational behavior be reliably distinguished from a human's — not because this question means the same thing as "can machines think," but because it is a question we can actually investigate, unlike the original.

This reframing matters enormously for how the standard objections should land. The complaint that the test "rewards deception" only bites if the test was supposed to detect an internal state (real understanding) using behavior as evidence, in which case a system could game the evidence without possessing the state. But if Turing's question was never about detecting an internal state — if the substituted question is the whole content of the inquiry, not a proxy for a separate, deeper fact — then "the machine passed by producing convincing behavior" is not a failure of the test to detect something real; convincing behavior of the specified kind simply is what the substituted question was asking about, and there is no further fact being missed. The objection assumes exactly the metaphysical ambition Turing's framing paper explicitly disclaims.

This is not to say Turing's move is beyond criticism — one can reasonably ask whether the substituted question is actually a good enough replacement for whatever cluster of concerns motivated the original one, or whether it quietly smuggles the metaphysical question back in through the audience's own interpretation of what "passing" would show. Most popular discussion of the Turing Test does exactly this: treating a machine's pass as evidence of genuine understanding, which is precisely the metaphysical reading Turing's own framing tried to head off. But this is a criticism of how the test has been received and mythologized, not of the narrower, more careful claim Turing actually made. The interesting philosophical move in the 1950 paper is not "indistinguishable behavior constitutes thought" — a strong and genuinely contestable metaphysical thesis — but "the question of thought is unanswerable

as stated, so let us ask a related, tractable question instead and be honest that we have changed the subject." Whether that substitution was a good one is a live and legitimate debate. But it is a different debate from the one seventy-five years of commentary has mostly been having, and Turing, reading his own paper carefully, seems to have anticipated that people would make exactly this mistake, and tried, not entirely successfully, to head it off in his opening page.

METAPHYSICS

Minds Without Selves

Some theories of mind treat the self as a unified controller behind experience. This essay defends the harder view: that unified selfhood might be a construction the mind produces, rather than the thing doing the producing.

Petra Lindgren Sankt Olof School, Stockholm

Ordinary self-understanding treats the self as something like a central controller: a unified subject that has experiences, makes decisions, and persists as the single owner of a stream of mental states. This essay defends a view, associated with Buddhist philosophical traditions and, independently, with certain strands of contemporary cognitive science, that this picture has the direction of explanation backwards — that the unified self is not the thing producing coherent experience and action, but a model the mind constructs, after the fact and for good functional reasons, out of processes that are not, at the level that actually generates behavior, unified at all.

The case for a controller-self usually rests on the sheer felt unity of experience: it does not feel like a committee of subsystems is negotiating my next action, it feels like a single "i" is deciding. But felt unity is exactly the kind of thing a good model can produce without being an accurate description of the underlying process — a modern automobile's dashboard presents a single, unified "engine status" indicator, and the driver experiences this as one coherent fact, even though the actual process underneath involves dozens of independent sensors and subsystems with no unified "engine self"

anywhere among them, only a well-designed summary display built for the driver's convenience. The feeling of unity, in other words, is evidence about what kind of summary the mind is presenting to itself, not directly evidence about what is doing the underlying work.

Split-brain research offers the clearest empirical support for taking this seriously as more than a philosophical possibility. When the hemispheres of the brain are surgically disconnected, each hemisphere can be shown, under controlled experimental conditions, to process information, form preferences, and initiate actions the other hemisphere has no access to and no awareness of — and each hemisphere, when tested independently, produces confident, seemingly unified reports of "its" reasons for acting, unaware that a separate process in the same skull is doing the same thing on the basis of entirely different information. Michael Gazzaniga's famous "interpreter" hypothesis proposes that a specific left-hemisphere system is specialized precisely for constructing a coherent, unified narrative explanation for behavior after the fact, regardless of whether that explanation accurately reflects the actual, possibly disunified causes of the behavior being explained.[1] If a dedicated interpreter system is producing our sense of unified selfhood as a narrative gloss on processes that are not, underneath, unified at all, then the ordinary feeling of being a single controlling self is not misleading in some minor, correctable way — it may be doing exactly what it evolved to do, constructing usable coherence, while being systematically uninformative about the actual, more fragmented architecture generating our behavior.

The obvious worry is that this proves too much, undermining the very notion of responsible agency that a unified self is usually invoked to ground: if there is no controller, who is making the decisions we hold people accountable for? But this worry assumes accountability requires a metaphysically unified controller, and it is

not clear this is true. A modern corporation is held legally and morally accountable for its actions without anyone thinking a unified corporate mind exists over and above its distributed departments, employees, and decision procedures; what grounds the accountability is a stable, trackable pattern of decision-making processes and their downstream effects, not a discovered metaphysical unity underlying them. A person's decision-making, even if it emerges from a genuinely disunified underlying architecture, can be tracked, predicted, and held responsible in the same way, as long as it produces a sufficiently stable pattern across time — which the narrative-constructing interpreter system, whatever its accuracy about underlying causes, reliably helps produce.

This is not a claim that the self is an illusion in the sense of being entirely without basis — the constructed narrative is doing real, useful, and remarkably stable work, coordinating behavior and producing the coherent patterns that accountability and planning depend on. It is a claim about what kind of thing is doing that work: not a discovered unity that experience merely reveals, but a genuinely constructed one, built by specific cognitive processes for good functional reasons, out of underlying material that was never, at the level where decisions actually get made, as unified as the finished construction feels from the inside.

NOTES

1. Michael S. Gazzaniga, *The Social Brain: Discovering the Networks of the Mind* (1985); the "interpreter" hypothesis is developed further across Gazzaniga's subsequent work on split-brain patients.

MASTHEAD

Minds and Machines — Issue 05

EDITOR-IN-CHIEF

Dominic Alvares

St. Xavier's Collegiate School, Kolkata

DEPUTY EDITOR, PHILOSOPHY OF MIND

Wen Jing Ho

Chinese International School, Hong Kong

FACULTY ADVISOR, POLITICAL PHILOSOPHY

Prof. Elena Marsh

Princeton University

WORKSHOP LEAD, PHILOSOPHY OF MIND

Dr. Julian Voss

National University of Singapore

REVIEWING EDITOR, EPISTEMOLOGY

Dr. Amit Sharma

University of Oxford