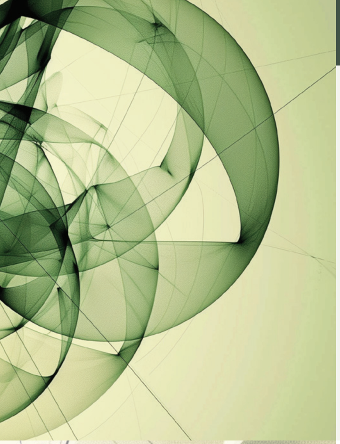


MINDFUL AI
FOUNDATION

Society of Minds:
A Philosophical Bridge to AI

Max Michaels and Rao Mikkilineni

Abstract

This paper advances Society of Minds as a philosophical bridge to AI: an operational account of how intelligent systems can become answerable to reasons, not merely capable of fluent outputs. Against scale-driven, monolithic approaches, we argue that normativity, explanation, and gov-ernance are not external “alignment layers,” but constitutive features of intelligence itself. Building on Minsky’s Society of Mind, we shift the locus of intelligence from a single model to an ensemble in dialogue—multiple minds, human and artificial, generating knowledge that is discernible in its reasons, criticizable in its claims, and extendable in its form.

We introduce a Mind–Brain–Body triad for the digital domain, defining Mind as the layer that compiles intent into bounded, auditable commit-ments and enforces critique-before-action. On this view, intelligence is not reducible to computation alone, but consists in normative control over information processes. Using Burgin’s General Theory of Information (GTI) as a formal bridge, we show how epistemic requirements (criticizability, provenance, revision) and dialogical requirements (justification, objection, legitimacy) can be translated into implementable constraints.

The framework operationalizes a familiar philosophical demand: that reasons, not outputs, must govern action. Philosophical constraints are rendered computationally explicit: memory becomes traceable, claims acquire provenance, and agency is routed through procedures of justifica-tion and refusal. Graph-structured memory grounds shared context; cog-nizing oracles enforce evidential warrant and the right to abstain; and lineage-aware, policy-governed knowledge structures preserve the conti-nuity of reasons across time. When interaction is treated as governed discourse, the characteristic failure modes of scale-first AI — opacity, brit-tleness, coherence debt, and normative slippage — become targets of de-sign rather than surprises of deployment. A management instantiation il-lustrates how governance kernels can convert generative outputs into policy-bounded commitments. Extending Minsky’s structural insight, we conclude that progress toward trustworthy AI requires not larger minds, but governable ones: a coalition of minds required to justify, remember, and defer before acting.

Keywords: Society of Minds; Mind; General Theory of Information (GTI); normativity; intentionality; criticizability; dialogical reasoning; philosophy of computation; Mindful Machine; AI safety

1. Introduction

Marvin Minsky’s Society of Mind [1] proposed an arresting architectural metaphor: a mind not as cathedral but as city—no singular nave of enlightenment, only neighborhoods of small, workmanlike agents whose traffic and trade give rise to thought. In this view, intelligence is civic, dis-tributed, a choreography of limited parts that, in aggregate, perform something like understanding. The brilliance of the metaphor outpaced the machinery of its era. Hand-tooled rules stood in for memory; brittle modules posed as judgment. The edifice persuaded as theory but creaked in practice, a scaffolding without the reinforced beams of learning or long-term coherence.

The landscape has shifted. Deep networks now ingest oceans of data, yet much of their power is opaque—black-box virtuosity without trans-parent reasons, fluency without common sense. Critics and champions alike converge on a common worry [2, 3, 4]. Large language models imi-tate the music of discourse while missing its meaning; they lack persistent memory, embodied world models, and the disciplined habit of explana-tion. Scale has delivered spectacle, not mindfulness [5]. If Minsky demysti-fied the mind by distributing it, today’s problem is the inverse: we have concentrated capability without distributing responsibility, reason, or re-call.

This paper answers that problem with a reframing: from a society within a mind to a Society of Minds—plural intelligences, human and arti-ficial, bound by norms of dialogue, shared memory, and the obligation to give reasons. The promise is not a bigger model but a better polity: sys-tems that argue, revise, and co-construct meaning; machines that are accountable to each other before they are persuasive to us. To orient the argument, we set out the paper’s goals:

- Recast the problem as philosophical, not merely technical: treat scale-only AI as a category mistake that confuses fluency with un-derstanding and probability with justification; propose Society of Minds as a route to intelligence that is answerable to reasons, not just capable of outputs.
- Specify a normative architecture of mind for the digital domain: advance the Mind–Brain–Body triad as an operational account of how agency can be made accountable (Figure 1): Body as situated sensing/acting, Brain as inferential and generative machinery, and Mind as the locus of intent, constraint, and critique that governs when action is permitted.
- Clarify what “Mind” contributes that models alone cannot: define Mind as the compiler of intent: the layer that turns goals into bounded commitments, evidence thresholds, and duties of absten-tion; in this view, memory and reasoning worth trusting are insepa-rable from provenance, revisability, and restraint (Figure 2; Table 2).
- Ground meaning and legitimacy in dialogue rather than prediction: treat intelligence as a practice of giving and asking for reasons, where claims gain standing through criticizability and contestation; interaction is not an interface feature but the medium through which knowledge becomes testable, corrigible, and ethically legible.

- Use GTI to bridge philosophy to implementable constraints: draw on Burgin's General Theory of Information to formalize epistemic and dialogical demands as requirements on information structures and processes (provenance, lineage, update rules, governance protocols), yielding a civic conception of intelligence that supports institutional learning without dissolving personal accountability.

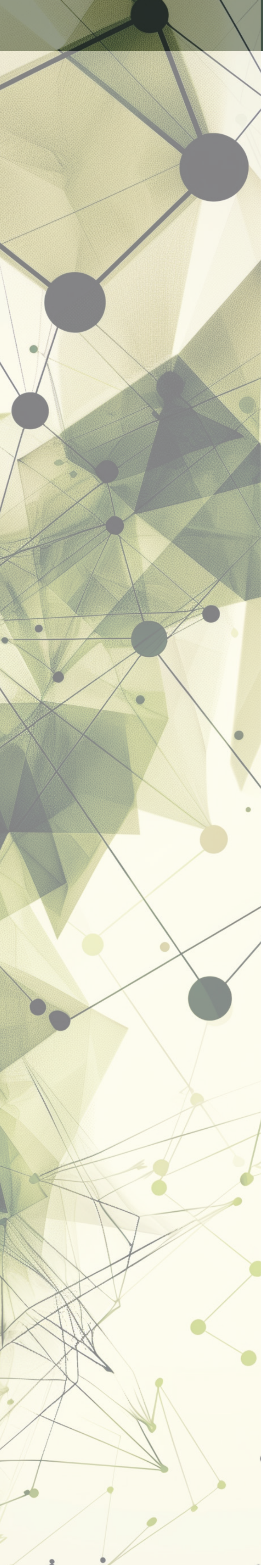
Stance and objections. We use "mind" in a functional-architectural sense rather than as a claim about phenomenal consciousness. Skeptical positions (e.g., Penrose-style non-computability or Chinese Room-style arguments) remain important philosophical constraints; our aim here is narrower: to specify an architecture in which explanations are criticizable, memory is lineage-aware, and action is procedurally governed, so that advanced AI systems can be more trustworthy in enterprise settings.

Overall, our approach represents a paradigm shift, not an incremental tweak. Rather than building ever-bigger black-box models, we propose a new blueprint for AI – one that emphasizes interaction, explanation, and governance among a community of minds. In the following sections, we detail this framework and its significance: first revisiting Minsky's original concept, then presenting the reimagined Society of Minds and its novel architectural components and finally discussing the broader implications for AI development.

2. Society of Mind Framework

Marvin Minsky's Society of Mind theory [1] remains a foundational idea in AI and cognitive science. Minsky argued that intelligence can emerge from numerous simple parts – a "society" of tiny agents inside one mind. In his view, each agent might be a minimal cognitive skill (a small rule, a heuristic, a feature-detector, etc.), and no single agent is smart on its own. Yet, when organized into the right hierarchies and feedback loops, their collective behavior produces what we recognize as thinking. As Minsky famously noted, "the power of intelligence stems from our vast diversity, not from any single, perfect principle" [1]. This perspective was a break-through in the 1980s: it shifted AI research away from searching for one grand algorithm and toward constructing intelligence via the interaction of many smaller processes.

However, the Society of Mind paradigm was very much a product of its time. Minsky's agents were largely conceived as hand-crafted symbolic programs – if-then rules, frames, or other modular routines. While con-ceptually elegant, such agents proved difficult to scale or adapt. Early at-tempts to build Society-of-Mind systems ran into coordination problems: without robust learning and memory, a collection of simple agents could just as easily produce chaos as coherence. By the 1990s, the field's focus had shifted toward two other approaches: on one hand, large neural net-works that learned patterns in a more monolithic way (trading interpret-ability for performance) [5], and on the other hand, dedicated expert sys-tems for narrow tasks [6]. Minsky's grand vision remained only partially realized; it lacked the detailed mechanisms to allow many components to learn together and form a truly integrated mind. In summary, the original



Society of Mind provided a powerful metaphor – intelligence from interaction – but implementing that metaphor required capabilities (like scalable learning, shared memory, resource optimization, and conflict resolution among agents) that were not yet available in Minsky's era.

Emergence is not enough. Large language models predict what is probable given training data; when they speak, they do not directly assert what is true, but what is statistically likely. Treating hallucination as noise to be washed away by more data misunderstands the structural problem: when internal operations are not constrained by reality, systems accu-mulate coherence debt—entropy produced by locally plausible outputs untethered from global factual alignment. Trusting emergence to smooth out coherence debt is like building a skyscraper without a blueprint and hoping physics will rescue the design.

Agentic AI and orchestration. Recent "agentic" approaches orches-trate LLM calls, tools, and sub-agents to complete multi-step tasks. This is a crucial direction, but without a legislating layer it can amplify failure modes: fluent agents can coordinate confidently around a shared error. Our contribution is to add a procedural governance layer that requires reasons, tracks provenance, and enforces constraints—so that orchestra-tion becomes accountable rather than merely effective.

Table 1 provides a high-level comparison of Minsky's Society of Mind versus our proposed Society of Minds, highlighting how the unit of analy-sis, nature of intelligence, and overall focus shift in the new paradigm.

3. Society of Minds Paradigm

The Society of Minds paradigm extends Minsky's original metaphor outward, from many agents within a single mind to many minds—human and artificial—interacting within a common ecosystem. Clarify here what "mind" means in this context: a coherent, semantically whole module with memory, reasoning, and explanation capabilities, not just a tool or sub-routine.

Unlike traditional multi-agent systems that focus on orchestration or task allocation, the Society of Minds introduces dialogical governance as its core innovation. Minds are not only required to act but to explain their reasoning to peers—human or artificial—and to accept critique or revi-sion. This avoids "post-hoc rationalization" by employing cognizing ora-cles and graph-structured memory to ground explanations in traceable reasoning chains [5].

Technically, this approach builds on prior paradigms without discarding them. Symbolic reasoning provides transparency, neural networks provide perception and fluency, and agentic controllers provide autonomy. The novelty lies in composition and communication: these diverse modules are knitted together through governance protocols (consensus algo-rithms, arbitration policies, conflict resolution strategies).

Practical demonstrations illustrate feasibility. In enterprise strategy, multiple AI agents (data analyst, scenario planner, ethics auditor) co-construct recommendations, each explaining its contribution. In med-ical diagnostics, specialist models (radiology, pathology, ethics) interact with human doctors, producing consensus diagnoses with documented reasoning. These principles are already being tested in video-on-demand and enterprise observability prototypes.

This design acknowledges new challenges: redundancy, potential groupthink, and coordination overhead. Computational cost is real, but histori-cal transitions—from single-core to multi-core, from monoliths to micro-services—show that efficiency follows once coordination mechanisms mature.

The concern that the digital genome might be too static or a single point of failure is addressed by noting that the genome is dynamic and evolving, updated by event-driven history and non-Markovian processes [9] should be understood as a policy-driven, evolving knowledge fabric rather than a fixed repository.
Thus, the Society of Minds paradigm is not just orchestration of modules but cultivation of relational intelligence: knowledge grows through ex-planation, critique, and revision across interacting minds. This dialogical orientation sets the stage for our discussion of epistemology, safety, and emergent coherence.

Medical Management operations as a proving ground. This work contains exactly the failure modes that defeat monolithic AI: ambiguous evidence, shifting constraints, and high cost of confident error. A Society of Minds implementation separates generation from justification: an LLM proposes options; a cognizing oracle demands assumptions; a critic chal-lenges coherence with market realities; the governance kernel enforces procedural constraints; and graph memory preserves why a decision was made so it can be revised when conditions change.

We return to this distinction in the Discussion (Figure 1), where the Body–Brain–Mind triad makes explicit why governance and critique must be treated as architectural layers, not afterthoughts.

Collective Mind, individual accountability. In this architecture, the Mind is not identified with a single expert. It is instantiated as the collec-tively encoded judgment of a community of practitioners, expressed through shared governance rules, critique protocols, and lineage-aware memory. Individual decisions remain the responsibility of specific human actors, who operate within—and contribute to—the evolution of this col-lective cognitive framework. This separation enables institutional learning without dissolving personal accountability.

From emergence to legislation: separation of powers. Biological sys-tems persist not merely because they are complex, but because they are legislated: a genome constrains repair, boundaries, and purpose. By anal-ogy, mindful machines require an internal separation of powers rather than a single monolithic network. We therefore specify three functional roles: (1) Worker (the LLM/Brain), a creative generator of possibilities; (2) Manager (the governance kernel/control plane), which holds procedural constraints over truth, purpose, and policy; and (3) Critic (the auditor), which evaluates the Worker's outputs against the Manager's constraints. In this architecture, "truth" is not a probability; it is adherence to a causal, legislative constraint.

The Society of Minds paradigm extends Minsky's metaphor outward, from many agents within a single mind to many minds—human and artificial—interacting within a common ecosystem. In this paper, "mind" denotes a criticizable, self-regulating capability that (i) produces discernible explanations for claims and actions, (ii) maintains lineage-aware memory (what is known, why it is believed, and where it fails), and (iii) governs action through procedural constraints. We do not treat "mind" here as a claim about phenomenal consciousness, but as an architectural function required for trustworthy intelligence.

Mind as the compiler of intent.

In Minsky's view, intelligence is an emergent "patchwork" of many small agents inside one mind, reflecting the symbolic AI era. In the Society of Minds, intelligence is an emergent coherence from many full minds interacting, reflecting a relational and dialogical approach.

	Society of Mind (Minsky)	**Society of Minds (Proposed)**
Unit of analysis	Small, simple agents within one mind	Full, autonomous minds (AI systems) in a community
Nature of intelligence	Emergent patchwork of micro-processes	Emergent coherence from re-lational interactions
Philosophy	Symbolic, reductionist (modular rules)	Holistic, relational (dialogue and diversity)
Era reflected	Early computing; hand-crafted AI mod-ules	Networked era; multi-agent systems and LLMs
Main focus	Simulating thought (no true self-awareness)	Cultivating understanding, coherence, and purpose
Key implication	"A mind emerges from many tiny agents."	"Meaning and understanding emerge when many minds in-teract."

A useful operational refinement, drawn from the "Compiler of Intent" framing, is to treat Mind as the layer that manufactures intent before generation and before execution: it translates high-level goals into bounded commitments—objectives, constraints, required evidence, and assurance criteria—that can be enforced and audited. In LLM-assisted systems, the abstraction level rises; what becomes scarce is not text gen-eration but the disciplined specification of "what we mean," under what constraints, and with what proof that the system stayed within bounds. In this sense, the model suggests; the system decides. [7]

Table 1 at the end of the manuscript present other definitions of "mind" in philosophy and AI.

What’s New?

The move toward a Society of Minds entails several novel conceptual shifts and technical innovations. In summary, our approach introduces the following key contributions to AI architecture and thought:

- Ethics Embedded in Architecture: We advocate moving beyond ex-ternal “after-the-fact” guardrails and instead baking in ethical rea-soning and constraints from the start. Concretely, we propose con-structs like Digital Genomes (governance kernel/control plane) and Cognizing Oracles to encode values and enable self-monitoring within AI minds [9]. A digital genome is a structured knowledge representation that carries not just facts but also the purpose and context of those facts – essentially encoding a system’s goals and constraints in its knowledge base [9]. Meanwhile, a cognizing oracle is a mechanism for transparency: it “looks inside” black-box models to extract human-readable explanations or detect doubt. By incor-porating these components, an AI mind can explain its reasoning and flag potential errors by itself, enforcing accountability internal-ly. In our Society of Minds, every claim should come with a rationale (via the oracles), and every mind’s knowledge is accompanied by provenance and usage context (via digital genomes). This built-in explainability and value-tagging means ethics is not an after-thought but part of the AI’s core design. (In practice, these ideas draw on recent work by Mikkilineni & colleagues [8], linking our ap-proach to emerging “mindful AI” prototypes [9]).

- Ethics as procedure, not doctrine. We do not propose a single moral theory “inside” the machine. Instead, ethical behavior is treated procedurally: the system must surface its policy basis, expose tradeoffs for critique, log provenance and accountability, and route ambiguous cases to human oversight. Ethics enters as governance signals, audits, and constraints—parameterized by enterprise policy, professional standards, and applicable law.

- Brain–Body–Mind Triad Model: What Is a Computer, and Is the Brain a Computer?[11] are questions that are fundamental to the evolu-tion of AI. We frame AI systems in a layered triad analogous to the relationship between body, brain, and mind in philosophical terms. The Brain represents the data-driven learners (neural networks, probabilistic models) that process inputs and produce candidate outputs. The Body represents the machine’s interface to the world – sensors, actuators, and also any fixed rule-based code (the deter-ministic machinery). The Mind is the overarching reasoning and regulating entity that interprets the Brain’s outputs, applies judg-ment, and decides actions by the machine. This triad offers a clear conceptual separation: the Body provides signals and actions, the Brain provides patterns and predictions, and the Mind provides meaning, goals and their execution. By explicitly modeling these layers, we can address current shortcomings: today’s AI “brains” (LLMs, etc.) have raw power but no self-aware mind to guide them. Our architecture adds that missing mindful layer on top of existing AI, providing what Kant might call the understanding and judgment to complement the mechanical processing of a brain. This also aligns with the idea that true intelligence requires integrating multiple perspectives – the physical (Body), the computational (Brain), and the rational/ethical (Mind).

- Hegelian Dialectic of AI Evolution: We interpret the trajectory of AI through a Hegelian lens of thesis, antithesis, and synthesis [12]. The thesis has been human cognition as we know it – limited by biology but rich in common sense and values. The antithesis is machine in-telligence as a brute-force simulation – enormously scalable and precise, yet lacking understanding and moral grounding. The syn-thesis, we argue, is a Society of Minds: a unification where human and machine minds interact in a shared system, combining their strengths while resolving their respective weaknesses. In practical terms, this means AI development should transition from humans and AIs working separately to humans and AIs working together in a principled way. Our framework explicitly enables such collabora-tion: human minds can be part of the society (e.g. a human expert might function as one "mind" in a problem-solving network), and the AI minds are designed to communicate in hu-man-understandable terms (through explanations and dialogues) rather than operate opaquely. This dialectical progression also un-derscores why a Society of Minds is timely: it addresses the current tension between the incredible capability of AI (antithesis) and the urgent need for control and meaning (thesis) by synthesizing a new paradigm where control is achieved through interaction and transparency.

- "Mindful Machines" – A New Evolved-State: We view Mindful Ma-chines as an achievable the goal of this paradigm. A Mindful Ma-chine is not characterized by sheer speed or the size of its model, but by qualities of memory, reasoning, self-regulation, and self-awareness of its limits. It is an AI that remembers context over long periods, that can reason through novel problems, that regu-lates its own behavior according to ethical principles, and that knows when to act, when to ask for help, or when to refrain from action. This stands in stark contrast to today's systems which are often "superficially fluent but clueless underneath" [5]. To achieve this, the Society of Minds design includes a graph-structured memory as a shared knowledge store, enabling persistent context and collective learning. All interactions between minds update this memory graph, which acts like an evolving knowledge base the so-ciety can draw on – thereby giving the machine a form of long-term memory and continuity that typical neural networks lack. Addition-ally, the multi-mind setup means the machine can deliberate: mul-tiple minds can debate a question or plan, yielding reasoning that is dialogical (many viewpoints considered) and discernible (steps can be traced). The result is an AI that behaves less like an "oracle" spit-ting out answers and more like a wise committee that can show its work. These Mindful Machines, we believe, would be safer and more reliable by design – they don't just produce answers, they produce explanations, justifications, and safeguards as part of their output. In essence, the Society of Minds approach shifts the metric of pro-gress from raw performance (e.g. benchmark scores) to coherent understanding and trustworthiness. Implications: So What?

- Why does this Society of Minds paradigm matter, and how we stand to benefit? We highlight several broad implications across conceptual, societal, and practical dimensions:

- Conceptual Leap: This framework shifts the AI discourse from purely optimizing performance metrics to fostering relational under-standing. It reframes AI progress not as a race for a single all-powerful model, but as the development of ecosystems of intel-ligent agents that learn to reason together. In doing so, it breaks the mold of thinking of intelligence as residing in one head (or one model), instead positioning intelligence as an emergent property of a network of interactions.

- Societal Value: By leveraging many minds, our approach offers a pathway to overcome the stagnation of human cognition in the face of accelerating machine capabilities. Rather than AI competing against humans, a Society of Minds is inherently a partnership model: it ensures that machine intelligence augments human intel-ligence, working with us in a transparent dialogue. This could help prevent scenarios where AI grows incomprehensible or uncontrolla-ble – instead, humans remain in the loop as part of the society, and the combined system can tackle problems neither could solve alone.

- Organizational Impact: Practically, the Society of Minds reframes AI as a strategic operating system for organizations. Decision-making in complex domains (enterprise strategy, scientific R&D, govern-ance) can be supported by a council of AI minds with different ex-pertise, much like a diverse team of advisors. This means decisions are informed by multiple perspectives (financial, ethical, technical, human impact) all generated in tandem by the AI society powered by Mindful Machines. The outcome is a more robust and well-rounded analysis, reducing single-point failures in corporate or policy choices. In short, AI becomes less of a monolithic tool and more of a distributed, dialogical process embedded in organiza-tional workflows.

- Ethics Reinvigorated: Our paradigm embeds ethical deliberation into the fabric of AI systems. Instead of relying solely on external compliance or post-hoc fixes, the Society of Minds has moral checks inherently through its multi-mind governance. This reframes the AI alignment problem: alignment is no longer just tuning a single model's behavior but designing a society in which unethical pro-posals are caught and corrected by other members. It mirrors how pluralistic democratic societies prevent extreme actions through debate and checks, rather than trusting any one actor blindly. Such an AI would be easier to audit and regulate, since it produces ex-planations and divides power among units. It aligns closely with emerging principles of ethical AI by design – moving from external oversight to internal conscience.

- Philosophical Underpinning: By drawing on philosophical insights (Kant's categories [13], Wittgenstein's language-games [14], Haber-mas's communicative rationality [20]), this framework grounds technical AI development in deeper theories of knowledge and meaning. Intelligence here is not just number-crunching; it's rela-tional, conversational, and normative. This helps bridge the gap between engineering and humanities in AI discourse. For instance, Wittgenstein argued that understanding is inherently social and contextual – our Society of Minds explicitly operationalizes that idea by requiring AI minds to negotiate meaning with each other and with humans. Habermas emphasized the importance of rea-son-giving in legitimate dialogue – our architecture ensures every decision can be questioned and explained. In effect, the paradigm connects AI's future to enduring human questions about how knowledge is created, shared, and validated.

- Towards a Sustainable Future: We define "Mindful Machines" as AI systems that are transparent, ethical, and collaborative by design – a stark contrast to the opaque, unyielding AI systems that domi-nate today. Embracing a Society of Minds vision steers AI develop-ment onto a more sustainable trajectory: one where progress is measured not just by capability, but by controllability and trust-worthiness. It offers a hopeful vision where increasing machine in-telligence does not equate to increasing risk, because greater capa-bility is coupled with greater internal governance and wisdom. As AI pioneer Gary Marcus has implored, we should demand "a better form of AI" – one that earns our trust by how it operates, not just by its results. The Society of Minds paradigm is a direct response to that demand, suggesting that the path to wise AI is through con-structing systems that inherently respect diverse perspectives, self-correct, and remain open to human guidance.

4. Discussion

The history of AI can be read as a succession of architectural experi-ments—each promising, each partial. Symbolic AI built transparent but brittle structures [16]. Deep learning scaled fluency and perception [16], but at the cost of opacity and reliability. Agentic systems intro-duced modularity and autonomy, but coordination remains fragile [16]. The Society of Minds does not discard these efforts but sublates them into a higher-order civic plan: modules, networks, and agents knit together through communication, critique, and governance. It is less about spectacle than about dialogue, less about monolithic per-formance than about relational coherence.

What is missing is a legislating layer that reduces coherence debt by forcing explanations, provenance, and critique before action.

Figure 1. In a Society of Minds, the "Mind" is distinguished from Brain and Body: Body closes the perception–action loop; Brain reasons and remembers; Mind governs action through constitutional intent and policy, enforcing explainability, abstention, and coherence.

- BODY (Sense / Act): generates information through embodied interaction; closes the perception–action loop.
- BRAIN (Reason / Remember): converts signals into representa-tions; performs inference; stores and organizes memory.
- MIND (Know / Reflect): governs reasoning and action using con-stitutional intent, legislation/policies, and self-regulation; enforces ex-plainability, abstention, and consistency requirements; manages co-herence debt.

The “Mind” is realized as the collectively encoded judgment of a community of practitioners, expressed through shared governance rules, critique mechanisms, and lineage-aware memory. Individual de-cisions remain the responsibility of specific human actors, who oper-ate within—and contribute to—the evolution of this collective cogni-tive framework. This separation enables institutional learning without dissolving personal accountability.

The implementation of the Society of Minds (outlined in Figure 2) embeds a control-theoretic view of body–brain–mind as a coupled loop that turns interaction-derived signals into structured, governed knowledge for action. In this view, the body provides sensing/acting that samples constraints in the world, the brain builds and updates representations (memory, models, inferences), and the mind performs meta-cognitive governance—applying a constitution (in-tent/teleonomy), self-regulation, and explainability constraints—so that the system’s knowledge remains extendible, its explanations re-main discernible, and its interventions remain reliable while minimiz-ing coherence debt over time.

Table 1. Operational Philosophy in Practice: Architecture element → medical assistant instantiation → risk addressed.

To clarify what we mean by a Mind layer, Figure 1 situates intelli-gence as a triadic structure rather than a single computational block. The Body executes and senses through machines and interfaces; the Brain reasons and remembers through models, planners, and memory systems; the Mind governs meaning by reflecting, critiquing, and con-textualizing action. The Society of Minds architecture explicitly in-stantiates this Mind layer as a collective, dialogical process rather than an emergent side effect of computation.

Operational philosophy. In high-stakes domains, the central fail-ure mode of generative AI is not a lack of fluency but a lack of disci-plined meaning: category errors, ungrounded claims, and the quiet drift of plausible text away from accountable truth. Operational phi-losophy treats epistemology as engineering. It translates commit-ments such as explainability, falsifiability, and responsibility into con-crete controls: provenance, critique loops, escalation thresholds, and audit trails. The Society of Minds architecture is best read in this key—to make knowledge criticizable before it becomes action.

Seen through the same lens, the Mind layer is best understood as an intent compiler: it turns philosophy into procedure—defining what counts as a valid claim, what must be verified, when to abstain, and when to escalate to accountable human authority [7].

From Competing Paradigms to Synthesis

The genealogy of AI is too often narrated as a battle of paradigms, each claiming sovereignty until its limits are exposed. Symbolic AI, with its clean hierarchies of rules, resembled a courthouse of log-ic—transparent, precise, but brittle against the mess of the real [10]. Deep learning, in turn, erected vast basilicas of statistical pattern, their stained glass glowing with synthetic images and fluent text, but with foundations obscured, inaccessible even to their makers [17]. The agentic turn added cloisters of autonomy, small chapels of modular function, yet coordination frayed as their ambitions grew [10]. A Soci-ety of Minds reframes these not as rival edifices to be abandoned, but as wings of a common city. Rules, networks, and agents each retain a role; their worth is preserved when stitched together by dialogue and governance. What once were contests of either–or are reimagined as both–and, with the binding mortar of explicit communication. In this plural architecture, breakthroughs no longer demand demolition. They can be grafted in, like new districts in an expanding city, without erasing the neighborhoods already built.

Explanations, Epistemology, and Emergent Safety

If knowledge is to endure, it must be tested in public, not whis-pered in closed chambers. A Society of Minds requires its agents—human and machine—to explain themselves, to give reasons that others may examine, contest, or refine. In this, the architecture mirrors Karl Popper's faith in falsifiability [18] and David Deutsch's in-sistence that progress is born of criticism [19]. Each exchange within the society is a miniature forum: conjecture offered, rebuttal posed, revision demanded. Knowledge is not locked in vectors of weight but inscribed in arguments that can be retraced and repaired. The scaf-folding of this discourse is technical as well as philosophical: graph-structured memory to anchor provenance, digital genomes to preserve lineage, governance protocols to enforce accountability. As Mikkilineni [5] observes, the contrast between biological and artificial systems is not merely one of scale but of structure. Biological minds regulate themselves, weave memory into lived experience, and expand knowledge through interaction; machines today, by contrast, remain pattern processors without autopoietic grounding or associative re-call. His proposal of "digital genome"–like structures, where knowledge is contextualized and updated through lineage and dia-logue, dovetails with the Society of Minds framework. It reinforces the point that knowledge must be both structured and social: not inert data, but discourse tested, criticized, and extended through interac-tion among minds—human and artificial alike. The design echoes Burgin's General Theory of Information [15], in which information is not inert record but living fabric—hierarchical, networked, open to growth. In this civic model, explanation is not ornament but require-ment. Safety emerges not from constraint alone but from conversa-tion: every claim must survive the critique of its peers. The model treats knowledge acquisition as more passive (receiving information) than proactive.

GTI as the Bridge between Epistemology and Dialogue

Burgin's General Theory of Information gives this paper its con-nective tissue: a way to treat philosophical demands as design con-straints. Kant's point—that understanding is shaped by rules—appears here as constraints encoded in the governance kernel. Popper and Deutsch supply the discipline of criticism: claims must travel with warrants, provenance, and a path to revision, or they do not earn the right to guide action. Wittgenstein keeps us honest about meaning, which lives in use; in practice, that means knowledge tagged by intent, context, and limits. Habermas adds legitimacy through discourse, op-erationalized as protocols that make objection admissible and escala-tion explicit. In GTI terms, information is not inert storage but a living fabric with lineage—updated through dialogue, tested through cri-tique, and stabilized through synthesis. Read this way, a Society of Minds is a regulated information ecology: mind is normative control over information processes, not a byproduct of scale.

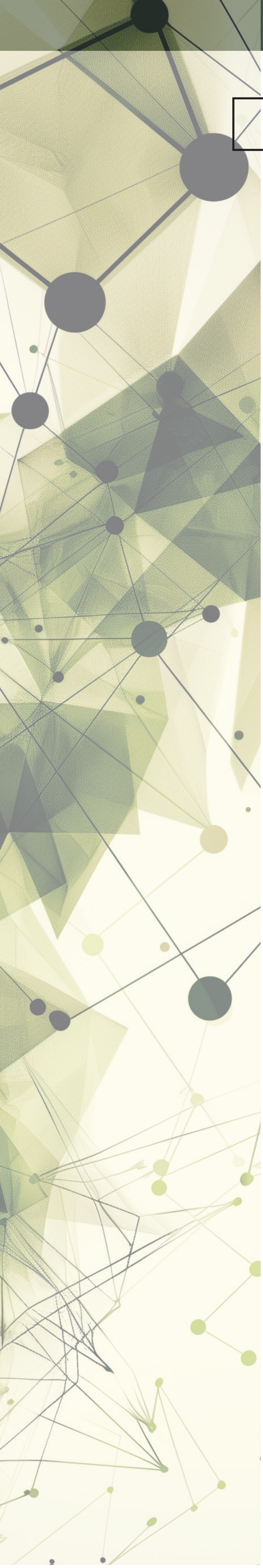
This bridge is made concrete in Table 2's architecture-to-risk mapping and in Figure 2's critique-before-action workflow, where norms become gates, and explanations become prerequisites for ex-ecution.

Toward a Mindful AI Paradigm

What emerges from this synthesis is not a refinement of existing tools but a new kind of polity. General intelligence may prove not to reside in a singular cathedral at all, but in the bustling agora between many smaller shrines of competence. The genius machine—solitary, oracular, self-sufficient—yields to a society whose resilience lies in plurality, whose wisdom arises from contention. In this vision, bench-marks are rewritten. Success is measured not only in narrow victories at task X, but in the ability to negotiate disagreement, to explain choices, to detect when an action strains against ethics or reason.

Our emphasis is on active reasoning, criticism, and agency could contrast with any passive models assumed in GTI discussions — which may assume that updating happens by exposure rather than by ques-tioning, choosing, or rejecting. Building blocks such as digital ge-nomes, oracles, and graph memory, when implemented in AI, can achieve autopoiesis and create associative memory – emulating hu-man thinking and understanding [11].

The Society of Minds is messy, perhaps, but like any city, its vitality lies in its intersections, its disputes, its capacity to absorb the new without unraveling the old. It is an architecture designed for growth, not perfection. To pursue this path is to accept that intelligence worth trusting will be born not of monologue but of dialogue, not of specta-cle but of sustained, self-correcting conversation.



5. Conclusions

Artificial intelligence stands at a crossroads. The last decade reward-ed scale and fluency. It also exposed a stubborn deficit: systems that can speak well still struggle to know when they should not speak, when they must justify, and when they must defer. We have built ma-chines of dazzling capability but shallow grounding—systems that simulate understanding while faltering at reasoning and explanation. The warnings of Marcus [2], LeCun [3], and Hinton [4] converge here: scale alone cannot deliver mindful intelligence. What is missing is not performance but normativity: the informational and computational structures by which claims become criticizable and actions become answerable. The Society of Minds offers an alternative: plural systems where minds critique minds, where explanations are demanded, where governance mirrors the checks and balances of human institu-tions.

In our view, mindfulness is not an emergent property of bigger mod-els. It is an architectural commitment, and a philosophical one. It re-quires a separation of powers—generation, governance, critique—and it requires interaction that produces reasons, not just outputs. That emphasis is not incidental to this paper; it is its wager: that mind-like intelligence is recognizable precisely where computation is con-strained by standards of justification.

Minsky insisted that mind runs through structure: many competences, stitched into a coalition that can explain, remember, and decide. Our thesis is that the next step runs through procedure: a governed Soci-ety of Minds, where cognizing oracles and traceable memory force justification and enable abstention, so action becomes accountable rather than merely persuasive. Burgin's General Theory of Information (GTI) supplies the bridge that Philosophies demands: it lets us treat epistemic and dialogical requirements not as aspirations but as im-plementable constraints—norms expressed as structures, and struc-tures expressed as governable information processes.

That is why the Mind–Brain–Body triad matters. The Body senses and acts through machines, interfaces, and workflows. The Brain reasons and remembers through models, planners, and graph memory. The Mind compiles intent into bounded commitments. It turns goals into constraints, evidence requirements, escalation rules, and audit trails. In short, the model suggests; the system decides.

Once intent is treated as compiled, governance stops being a bolt-on. It becomes a control plane, or governance kernel, that routes action according to risk. Medication changes, diagnosis claims, and triage decisions do not belong to a single probabilistic stream. They belong behind gates: provenance checks, oracle-based warrant, dissent han-dling, and named accountability for release. In this framing, account-ability is instantiated at the point of action, while Mind is distributed across the collective judgment of the practitioner community that uses and improves the system over time.

This is the practical meaning of a Society of Minds. Intelligence be-comes civic. Multiple minds, human and artificial, argue their way to-ward a decision. Graph-structured memory preserves the trail of rea-sons. Cognizing oracles enforce a discipline of evidence and absten-tion. Digital genomes provide lineage-aware policy and context so knowledge can be updated without erasing where it came from. The aim is not merely better answers, but better epistemic behavior.

Popper and Deutsch remind us that knowledge advances through criticism [18,19]. Our architecture makes criticism procedural. Hegel [12] gives the name for the motion: not compromise, but synthesis that preserves what works while cancelling what fails. In this archi-tecture, symbolic precision, neural breadth, and agentic sequencing interlock rather than collide. They become components whose incen-tives and interfaces must be designed to reduce conflict and increase coordination—so the system can earn trust by design rather than demand it by authority. Hence the philosophical payoff is operational, not ornamental.

The implications of the Society of Minds paradigm

This paradigm does not promise perfection. It will err and quarrel, but unlike a solitary black box, it contains the means of its own cor-rection. Its resilience lies in plurality, its wisdom in dialogue. Early prototypes—from enterprise decision-making to clinical set-tings—suggest feasibility, while familiar histories in computation (multi-core systems, microservices, distributed consensus) remind us that complexity, once feared, can be governed.

The challenge ahead is to formalize governance protocols, devel-op metrics for dialogical intelligence, and establish efficiency bench-marks that make critique scalable without making it performative. Yet humanity's greatest institutions—science, law, democra-cy—emerged not from suppressing plurality but from cultivating it under rules of evidence and argument. If AI is to become worthy of re-liance, it will need analogous institutions in silicon.

To embrace a Society of Minds is to renounce the fantasy of a sol-itary genius machine and instead imagine intelligence as civic, plural, and dialogical. Technically, it means abandoning the monolith in fa-vor of a federation: modular intelligences, interoperable by design, each adding perspective to deliberation. Governance, too, acquires a new contour. Inside the system, governance means embedding checks and balances—minds that audit, minds that veto, minds that explain. Outside, it means systems that welcome scrutiny because their internal discourse produces explanations before it delivers conclusions. A Society of Minds, by design, acknowledges fallibility and subjects it-self to critique. It is a humility machine, not an oracle.

Philosophically, this paradigm honors old insights with new ar-chitecture. Kant taught us that understanding is not given but im-posed [13], Wittgenstein that meaning is forged in use [14], Habermas that legitimacy lies in the force of the better argument [20]. The Soci-ety of Minds operationalizes these claims:

knowledge must be testa-ble, reasons must be shareable, and dialogue must be the medium of progress. This is intelligence not as spectacle, but as a disciplined practice of justification.

Building such systems will require patience and pluralism in our own institutions. Researchers must resist crowning a single paradigm as sovereign. Companies must learn to distribute responsibility across interoperable modules rather than hoard power in proprietary for-tresses. Policymakers must regulate for transparency and criticizabil-ity, not simply for throughput. The work is technical and cultural: to create an ecosystem that mirrors the society of minds we seek to en-gineer.

The road ahead is demanding but legible. We need shared-memory protocols, inspectable governance policies, metrics for dialogical performance, and empirical validation in high-stakes settings. Yet the direction is clear. If scale gave us remarkable speech, structure must give us accountable action.

We close with a call. Let us build not just faster minds, but wiser ones—minds in dialogue with each other, with humanity, and with the sustainable future we seek to create.

The Choice Before Us

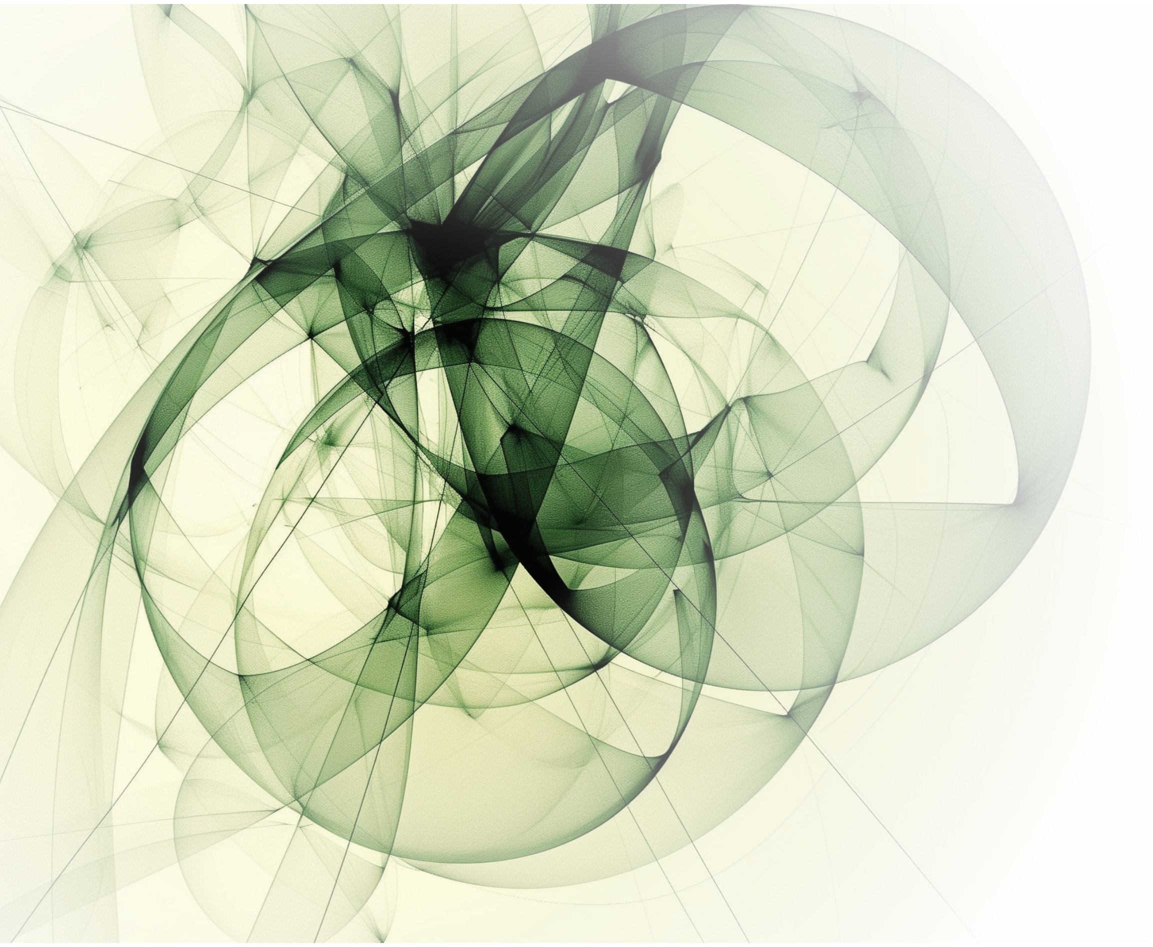
The history of intelligence, natural or artificial, is less a tale of solitary genius than of communities that learn by exchange. Human civilization advanced because minds argued, collaborated, and cor-rected each other's errors, turning private judgment into public rea-sons. To build machines that mimic intelligence without this social discipline is to produce systems that are fluent yet unanswera-ble—capable of outputs, poor at accountability. The Society of Minds paradigm begins from the opposite premise: intelligence matures through dialogue, and reliability through critique.

On this view, a mindful machine is not defined by confidence, but by conduct under criticism. A system becomes mind-like precisely when it cannot merely emit an answer, but must expose its reasons to challenge—internally, among specialized components with distinct roles, and externally, to humans who can contest, revise, and withhold consent.

The architectural move is simple to name and difficult to realize: legislate. Separate the Worker that generates from the Manager that constrains, and require a Critic that audits. This separation of powers will not eliminate error, but it can prevent error from acquiring au-thority. Put plainly, the Mind is the compiler of intent: it specifies the constraints, evidence thresholds, and assurance required for action, so generation is procedurally governed rather than merely persuasive [7].

In enterprise strategy, where decisions are made under uncer-tainty and later judged by consequences, this architecture makes learning cumulative. Decisions become revisable because their rea-sons are preserved. Knowledge becomes extendable because it carries provenance and context. A Society of Minds does not promise perfec-tion; it will err and disagree. But unlike a solitary black box, it contains the means of its own correction: plurality as resilience, dialogue as discipline.

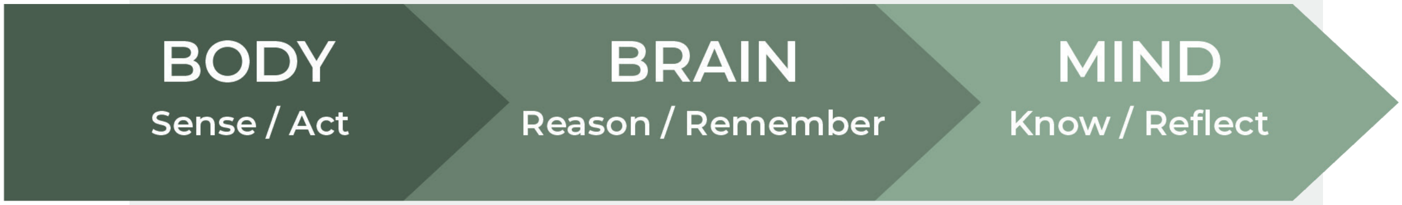
The future worth building is not an all-knowing oracle that speaks and expects belief, but a civic intelligence that earns trust the old-fashioned way: by giving reasons, keeping memory, and changing its mind when the evidence changes. The task before us is to build systems that deliberate, explain, and submit to critique—machines in dialogue with one another, with us, and with the future we are busy designing.



References:

1. Minsky, M. *The Society of Mind*; Simon & Schuster: New York, NY, USA, 1986.
2. Marcus, G. Deep Learning: A Critical Appraisal. *arXiv* 2018, arXiv:1801.00631. Available online: https://arxiv.org/abs/1801.00631 (accessed on 24 September 2025).
3. LeCun, Y. A Path Towards Autonomous Machine Intelligence. *Open Review* 2022. Available online: https://openreview.net/forum?id=BZ5a1r-kVsf (accessed on 24 September 2025).
4. Hinton, G. The Future of Artificial Intelligence: Opportunities and Risks. *Philosophical Transactions of the Royal So-ciety A* 2023, 381, 20220049. https://doi.org/10.1098/rsta.2022.0049.
5. Mikkilineni, R. General Theory of Information and Mindful Machines. *Proceedings* 2025, 126(1), 3. https://doi.org/10.3390/proceedings2025126003.
6. Newell, A.; Simon, H.A. GPS, a Program that Simulates Human Thought. In *Computers and Thought*; Feigenbaum, E., Feldman, J., Eds.; McGraw-Hill: New York, NY, USA, 1963; pp. 279–293.
7. Mikkilineni, R. The Age of the Compiler of Intent: Why LLM-Assisted Development Raises the Abstraction Level—and Why Governance Becomes the Differentiator. Working Paper, Version 2.1; 2026.
8. Mikkilineni, R. Mark Burgin’s Legacy: The General Theory of Information, the Digital Genome, and the Future of Machine Intelligence. *Philosophies* 2023, 8, 107. https://doi.org/10.3390/philosophies8060107.
9. Kelly, W.P.; Coccaro, F.; Mikkilineni, R. General Theory of Information, Digital Genome, Large Language Models, and Medical Knowledge-Driven Digital Assistant. *Computer Sciences & Mathematics Forum* 2023, 8(1), 70. https://doi.org/10.3390/cmsf2023008070.
10. OpenAI. GPT-4 Technical Report. *arXiv* 2023, arXiv:2303.08774. Available online: https://arxiv.org/abs/2303.08774 (accessed on 24 September 2025).
11. TFPI. What Is a Computer, and Is the Brain a Computer? 2024. Available online: https://tfpis.com/2024/04/21/what-is-a-computer-and-is-the-brain-a-computer/ (accessed on 24 September 2025).
12. Hegel, G.W.F. *Phenomenology of Spirit*; Oxford University Press: Oxford, UK, 1977 (orig. 1807).
13. Kant, I. *Critique of Pure Reason*; Cambridge University Press: Cambridge, UK, 1998 (orig. 1781).
14. Wittgenstein, L. Philosophical Investigations; Blackwell: Oxford, UK, 1953.
15. Burgin, M. *Theory of Information: Fundamentality, Diversity and Unification*; World Scientific: Singapore, 2010. https://doi.org/10.1142/7438.
16. Burgin, M.; Mikkilineni, R. On the Autopoietic and Cognitive Behavior of Digital Genomes (governance kernel/control plane). *EasyChair Preprint* No. 6261, 2021. Available online: https://easychair.org/preprints/preprint_down
17. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* 1986, 323, 533–536. https://doi.org/10.1038/323533a0.
18. Popper, K. *Conjectures and Refutations: The Growth of Scientific Knowledge*; Routledge: London, UK, 1963.
19. Deutsch, D. *The Fabric of Reality*; Penguin: London, UK, 1997.
20. Habermas, J. *The Theory of Communicative Action*, Volume 1; Beacon Press: Boston, MA, USA, 1984.

Figure 1. From Computation to Reflection: The Body–Brain–Mind Triad in Mindful Machines. The Body layer senses and acts, the Brain layer reasons and remembers, and the Mind layer reflects, critiques, and contextualizes decisions. A Society of Minds realizes this Mind layer as a governed, collective process that renders intelligent behavior explainable, contestable, and accountable.



- BODY (Sense / Act): generates information through embodied interaction; closes the percep-tion–action loop.
- BRAIN (Reason / Remember): converts signals to representations; performs inference; stores/organizes memory.
- MIND (Know / Reflect): governs reasoning and action using constitutional intent, legisla-tion/policies, self-regulation; enforces explainability/ abstention/consistency requirements; manages coherence debt.

In a Society of Minds, the "Mind" is realized as the collectively encoded judgment of a community of practitioners, expressed through shared governance rules, critique mechanisms, and lineage-aware memory. Individual decisions remain the responsibility of specific human actors, who operate within—and contribute to—the evolution of this collective cognitive framework. This separation enables institutional learning without dissolving personal accountability.

Figure 2. Society of Minds in operation: intent-compiled, governed, critique-before-action workflow for a Mindful Machine. The Worker (LLM/Brain) proposes; the Manager (Mind/governance kernel) compiles intent into constraints and evidence gates; the Critic audits warrant, provenance, and uncertainty before the Body executes. This separation of powers converts probabilistic generation into accountable, traceable action.

This figure illustrates a Mindful Machine implemented as a "Society of Minds" workflow, designed to neutralize the risks of generative AI in medicine by replacing a single black-box model with a system of checks and balances. At the center is a collaborative loop: the Worker generates clinical drafts, the Manager constrains them using hard safety rules and patient context, and the Critic audits the result for contradictions, missing evidence, or hallucinations. The cycle is governed by a Control Plane grounded in clinical guidelines and supported by Cognizing Oracles that require explicit warrant and provenance for each claim. Crucially, the process terminates in accountable human agency: the Attending Physician remains the named authority who must approve any critical action such as triage escalation, medication changes, or diagnostic claims.

Figure 2 operationalizes the mapping in Table 2 by showing how each architecture element instantiates concrete system roles and directly mitigates specific risk classes.

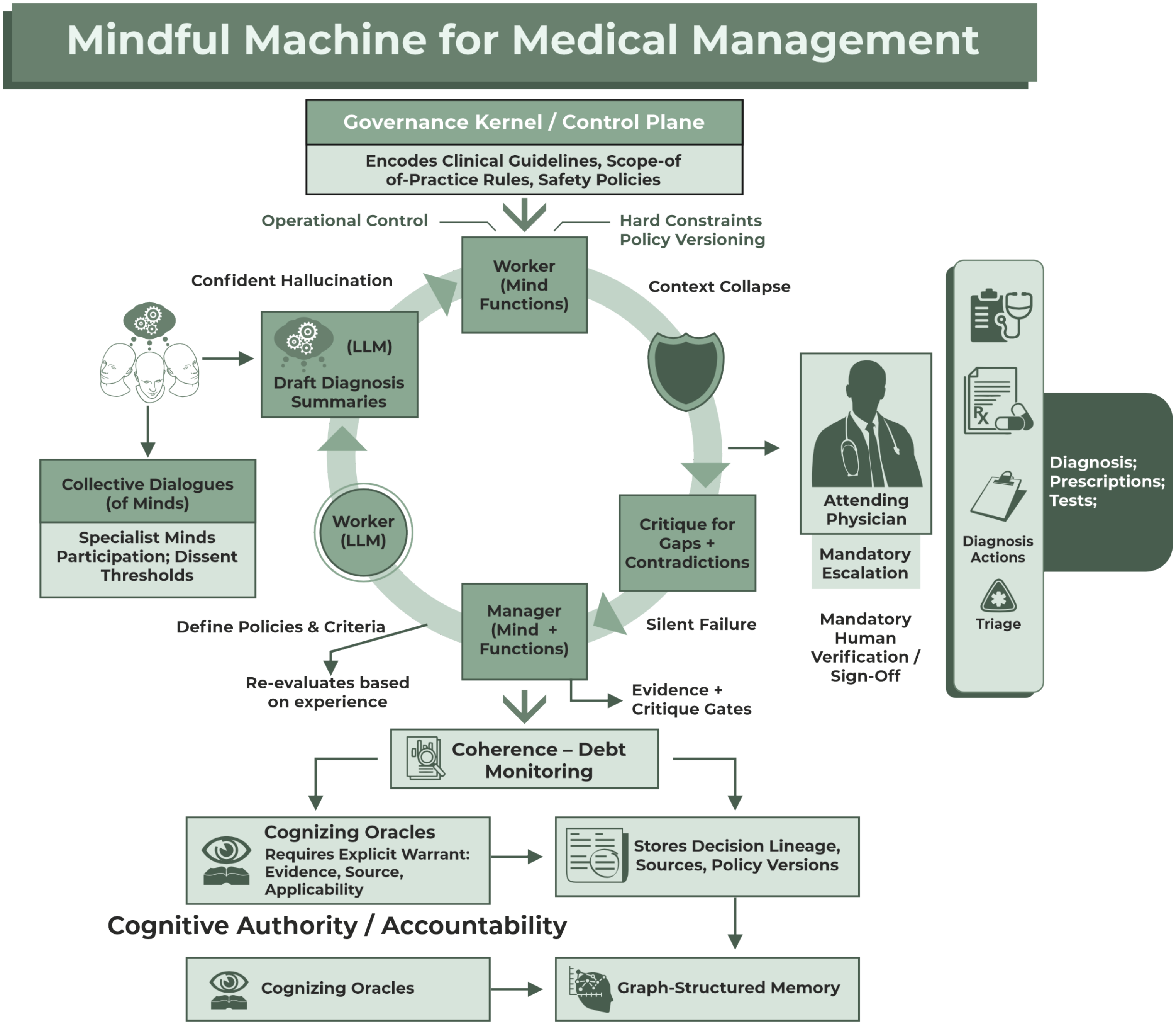

Mindful Machine for Medical Management
Governance Kernel / Control Plane
Encodes Clinical Guidelines, Scope-of of-Practice Rules, Safety Policies
Operational Control
Hard Constraints
Policy Versioning
Worker (Mind Functions)
Confident Hallucination
Context Collapse
(LLM)
Draft Diagnosis Summaries
Collective Dialogues (of Minds)
Specialist Minds Participation; Dissent Thresholds
Worker (LLM)
Critique for Gaps + Contradictions
Attending Physician
Mandatory Escalation
Mandatory Human Verification / Sign-Off
Diagnosis Actions
Triage
Diagnosis; Prescriptions; Tests;
Manager (Mind + Functions)
Define Policies & Criteria
Silent Failure
Re-evaluates based on experience
Evidence + Critique Gates
Coherence – Debt Monitoring
Cognizing Oracles
Requires Explicit Warrant: Evidence, Source, Applicability
Stores Decision Lineage, Sources, Policy Versions
Cognitive Authority / Accountability
Cognizing Oracles
Graph-Structured Memory

Table 1.
Working definitions of "mind" in philosophy and AI

Thinker / tradition	Core definition / emphasis	Implication for AI	Relation to Society of Minds
Kant (epistemic)	Mind organizes experience into intelligible structure	Intelligence needs organizing principles, not just data	Implements organizing principles as governance kernel + criticizable explanations
Minsky	Mind emerges from interacting internal agents (a "society")	Decompose cognition into cooperating parts	Extends to many whole minds in dialogue, with shared memory and governance
Penrose	Understanding may be non-computable; mind not reducible to algorithm	Skepticism about Strong AI claims	Brackets consciousness; focuses on architected, criticizable cognition and accountability
LeCun	World models, memory, planning; self-supervised learning	Need persistent state and planning beyond next-token prediction	Adds dialogical critique, lineage, and procedural governance over modules
Hinton	Emergent representations in neural nets; richness and redundancy	Scaling yields competence but risks opacity	Adds governance kernel/control plane to reduce coherence debt and improve auditability
Hassabis	Integration of perception, memory, imagination, planning	Build multi-component systems that simulate and plan	Adds civic dialogue and internal checks (critic/auditor roles)
Michaels & Mikkilineni	Mind = criticizable explanation + lineage-aware memory + procedural self-regulation	Design Mind as a governing layer over Brain & Body	Society of Minds is the organizational form that makes this scalable and accountable

Table 2.
Society of Minds as Operational Philosophy: Architecture element medical assistant instantiation

Architecture element	Medical assistant instantiation (real system)	Risk addressed	Operational control	Cognitive authority / Accountability
Governance kernel / control plane	Encodes clinical guidelines, scope-of-practice rules, safety policies; versioned policy pack	Unsafe or non-compliant recommendations; guideline drift	Hard constraints; policy versioning; audit logs	Collective physician community defines policies; attending physician accountable for action
Worker (LLM / Brain)	Generates candidate differentials, summaries, and draft explanations	Confident hallucination; automation bias	Outputs must pass evidence + critique gates; abstain on insufficient evidence	Collective standards shape gates; attending physician approves use
Manager (Mind functions)	Constrains Worker outputs via governance kernel + patient/context state	Context collapse; inappropriate generalization	Context checks; rule-based exclusions; required assumptions list	Collective protocols; attending physician signs off
Critic (Auditor)	Evaluates Worker–Manager dialogue for evidence gaps and contradictions	Silent failure; overconfidence	Contradiction detection; uncertainty thresholds; mandatory escalation	Collective thresholds; attending physician resolves
Cognizing oracles	Require explicit warrant: evidence, source, and applicability before any claim is displayed	Unverifiable diagnosis claims; spurious rationale	Explain-or-abstain; provenance required; confidence calibration	Collective criteria; attending physician responsible
Graph-structured memory + provenance	Stores decision lineage: sources, rationale, policy versions, dissent, outcomes	Non-auditable reasoning; irreproducible decisions	Traceable reasoning; post-hoc audit; update triggers	Collective learning; attending physician accountable per case
Always-verified actions	Medication changes, diagnosis claims, and triage are gated to physician review	Patient harm from over-automation	Mandatory human verification; escalation workflow	Attending physician
Coherence-debt monitoring	Flags mismatch across evidence, policy constraints, and outputs over time	Locally plausible but globally wrong outputs	Drift detection; periodic re-validation; rollback to stable policies	Collective governance group curates; attending physician remains accountable
Collective dialogue (Society of Minds)	Specialist minds (radiology/pathology) + human clinicians participate; dissent captured	Single-point-of-failure decisions; groupthink	Consensus with dissent logging; minority report prompts	Collective community is cognitive authority; attending physician accountable