MINDFUL AI
FOUNDATION

Beyond the World-Model:
Why Human-Level AI Needs More Than Physics

Max Michaels | Kenzo Fujisue | Rao Mikkilineni, Ph.D

The LLM era has unfolded with Cambrian exuberance and renaissance-scale participation, making even its excesses feel generative. Seemingly overnight, machines began to write, reason, translate, code, tutor, and converse with a fluency that unsettled cognitive and cultural boundaries between machine and human output. The distinctions, between tool and collaborator, no longer felt entirely secure.

But such spectacle invites correction, and that correction has come from Yann LeCun, perhaps the most credible critic of language-model triumphalism. His argument is serious and, in one important respect, convincing: if we want human-level intelligence, language alone will not suffice. We need systems that learn from the world itself, through perception, action, consequence, and prediction. In that sense, world models offer a real advance. They promise a stronger account of causal representation and a more plausible path for robotics than language-only systems can provide. They ask the machine to confront the grain of reality, not merely its linguistic trace.

Yet every corrective carries the risk of becoming its own overstatement. LeCun's world-model program is a necessary turn away from the fantasy that next-token prediction is the whole of thought. It rightly reminds us that text is not the whole of reality, and that eloquence, however, is not the same thing as grounded intelligence. But what is compelling within the domain of embodied prediction is now being asked to support a much larger claim: a theory of human-level intelligence itself.

That is where the more difficult question begins. Not which system wins, but what kind of intelligence each architecture is suited to produce. LLMs excel where intelligence moves through language, explanation, retrieval, and dialogue. World models appear stronger where intelligence must anticipate consequences in dynamic physical environments before acting. But the next frontier may lie elsewhere: in systems that remain coherent across memory, governance, value conflict, and long-range human consequence. What follows, then, is not a dismissal of world models but an argument about scope. World models are essential, but partial. The deeper task is not merely to replace language with pixels, or to build better machinery for external prediction, but to cultivate systems that can enter human life without flattening it, systems rich in memory, alive to relationship, formed in language, answerable to ethics, and capable not only of simulating futures, but of discerning which futures deserve a place in the world we share. That missing governing layer, spanning memory, purpose, interpretation, and value-sensitive action, is what we call **Mindful AI**. What follows is a discussion about the scope of World-model in our quest for human-level intelligence.

“If you are interested in Human-level AI, don’t work on LLMs.”

Some provocations merely inflame; the better ones illuminate. LeCun’s belongs to the second kind. It usefully punctures the mythology that scaling language models alone will inevitably yield full-spectrum intelligence. LLMs are not complete minds. They do not natively possess persistent memory, grounded agency, robust self-modeling, or reliable long-horizon planning in the physical world.

But the declaration also overreaches. Language is not a decorative layer added late to intelligence. It is the medium through which human beings coordinate action, preserve memory, build institutions, transmit culture, negotiate norms, and repair one another. If the next era of AI is to be lived not only in factories and vehicles but in classrooms, clinics, offices, homes, courts, and conversations, then language is not a distraction from intelligence. It is one of its highest expressions.

The issue is not world models versus LLMs. It is that the two illuminate different design problems. World models address embodied prediction, causal structure, and action in physical environments. LLMs, for all their incompleteness, have shown that human-level intelligence must also move through culture, language, memory, interpretation, and collaboration. A theory of intelligence that neglects this second domain will remain partial, because human beings do not live by physics alone. We live in worlds saturated with meaning.

“World Model: A path towards autonomous machine intelligence.”

That phrase contains the first boundary line. The operative word is autonomous. Autonomy is a real engineering achievement. A robot, vehicle, or planning system must perceive, infer, predict, and act without continuous supervision. In that setting, a world model makes deep sense. The machine must estimate hidden state, anticipate consequences, and choose among possible actions. It must become skillful in the grammar of environments.

But autonomy is not the only horizon for AI, and it may not even be the most consequential one. The systems entering daily life are not only drivers, drones, or domestic robots. They are assistants, tutors, editors, research partners, memory prosthetics, mediators, and increasingly the interface through which people navigate institutions and one another. Their decisive challenge is not merely autonomous action. It is competent presence in human contexts. A path toward autonomous machine intelligence may be a path toward better robotics. It does not, by itself, settle the larger question of humane machine intelligence.

“Any house cat can plan highly complex actions.”

A well-aimed provocation can clear the air. LeCun’s does exactly that. The point is easy to grant. A cat can navigate space, infer affordances, and perform embodied feats that no LLM can approach. The example is useful because it reminds us of Moravec’s paradox: tasks that are easy for organisms are often extraordinarily hard for machines.

But examples can reveal a blind spot as well as a truth. A cat is excellent at being a cat. A language model is not. Fair enough. But the human worlds now being transformed by AI are not organized chiefly around stalking prey, clearing tables, or balancing on ledges. They are organized around language, institutions, norms, narratives, memory, interpretation, and care.

A cat cannot explain a medical report to a frightened patient. It cannot help a student understand Kant. It cannot preserve the continuity of a research project across months of conversation. It cannot mediate conflict between colleagues or remember the emotional stakes of a family decision. These are not side functions of civilization. They are central ones. The cat example proves something real, but narrower than it first appears. It shows that current language systems are not sufficient for embodied competence. It does not prove that cat-like competence is the master key to all forms of intelligence that matter in civilization. Physics is necessary for a robot. It is not sufficient for a companion.

“Our world model needs to be trained from sensory inputs.”

We have all internalized that a picture is worth a thousand words. In one sense, LeCun’s program begins there: intelligence must learn from the world as seen, touched, and acted upon, not merely from its verbal residue.

Human beings do not grow up on text alone. We learn from seeing, touching, moving, grasping, colliding, trying, failing. A child knows something about gravity before grammar, about objects before propositions, about persistence before syntax. Any serious theory of embodied intelligence must reckon with that. This is the conceptual core of the program.

But the existence of sensory grounding does not settle the larger question before us. The systems now reshaping education, law, software, medicine, writing, administration, and research are not primarily failing because they never watched enough balls roll off tables. They are failing because they forget too much, flatten context, lack durable self-models, do not reliably track values across time, and still struggle with the moral and emotional texture of human situations.

The gap is not only between text and sensation. It is between prediction and meaning. A machine can learn the dynamics of objects and still know very little about obligation, grief, fairness, dignity, irony, or tenderness. It may model the world and yet remain estranged from the human uses of the world.
Both GPTs and autonomous driving began, in a fundamental sense, not from raw innocence but from inheritance. GPTs were trained on vast archives of books, websites, code, and conversation, the sediment of culture already written down. Autonomous driving did not begin by rediscovering roads from first principles, but by drawing on maps, traffic laws, lane markings, driving conventions, sensor datasets, and millions of miles of recorded human behavior. In both cases, progress came not from forcing the machine to reinvent civilization from scratch, but from giving it access to civilization’s accumulated traces. Intelligence, artificial or human, rarely begins in a vacuum. It begins in a world already structured by memory.

One unresolved question in the world-model program is whether intelligence must earn its legitimacy by rediscovering from sensory noise what civilization already knows in explanatory form. Human beings do not begin each generation by relearning gravity from scratch. They inherit laws, models, tools, and cultural practices. A serious architecture for human-level AI may need to explain not only how machines learn from the world, but how they inherit from it.

The world-model is often positioned as though a picture were worth not only a thousand words, but also a theory of mind, a moral vocabulary, and a social constitution. It is not. Images can ground prediction. They do not, by themselves, ground judgment.

“We are using hierarchical JEPA to build universal action-conditioned causal models of any complex system.”

This is where the ambition of the program becomes both most impressive and most vulnerable. LeCun’s Joint-Embedding Predictive Architecture does not attempt to reconstruct the world in all its sensory detail. Its wager is that intelligence requires prediction in representation space: future abstract states, not raw pixels. The gain is efficiency and causal abstraction. Predictive deep learning avoids the noise problem that haunts generative reconstruction by forecasting latent states rather than every visual contingency.

LeCun’s Joint-Embedding Predictive Architecture does not attempt to reconstruct the world in all its sensory detail. Its wager is that intelligence requires prediction in representation space: future abstract states, not raw pixels. The open question is whether such latent simulation, however powerful, is enough for memory-bearing, norm-sensitive participation in human worlds.

The phrase “any complex system” suggests an architecture broad enough to scale from perception to causality to action, perhaps even from particles to organisms to societies. Yet a society is not merely a more complicated pile of physics. It is not just another dynamical system awaiting compression into latent causal structure. Human life is shaped by interpretation, normativity, memory, symbolism, ritual, power, promise, taboo, grief, aspiration, and story.

The danger here is not technical ambition. The danger is ontological flattening. When the framework stretches from physical causality toward human systems, it risks treating what matters most about a society as if it were simply that it can be modeled, predicted, and acted upon. But what matters most about a society may be that it must be understood with restraint.

The deeper difference between the two programs is not merely technical but architectural. LeCun’s world model is organized around predictive competence: learning abstract latent representations that permit causal forecasting and action planning. The Mindful AI alternative begins from a different concern: not only whether the system can predict the next state of the world, but whether it can preserve coherence across memory, purpose, interpretation, and value. One architecture minimizes prediction error. The other asks how intelligence remains answerable to what it must carry forward.

A system may be excellent at minimizing prediction error and still poor at managing what might be called coherence debt: the gap between immediate success and long-range consistency with memory, purpose, and human norms. World models do not yet answer for institutional continuity, for coherence across long-lived roles and commitments, or for the kind of coherence debt that accumulates when memory, explanation, and governance are bolted onto a system as afterthoughts. The world model can imagine futures in latent space. The harder problem is learning which futures are worth bringing into human life.

“All of our interactions with the digital world will be mediated by AI assistants.”

Here the argument turns. The earlier emphasis is on cats, robots, intuitive physics, sensory inputs, planning, and action-conditioned causal modeling. But now the future arrives in another form: not the autonomous robot in a room, but the AI assistant mediating digital life.

Once AI is framed as assistant rather than agent, the center of the design problem changes. The challenge is no longer only competent action in an external environment. It is competent presence in a relational, linguistic, and normative environment. An assistant must remember, explain, adapt, defer, calibrate, contextualize, and sustain continuity. It must remain sensitive to role, vulnerability, history, and institutional context.

That means the decisive missing architecture is not only a world model. It is a model of the human situation: a representation of context, memory, values, uncertainty, relationship, and constraint. Human-facing AI fails less often because it cannot predict the next state of the world than because it cannot manage coherence across time, role, value conflict, and institutional context.

“World model will constitute a repository of all human knowledge and culture.”

This is the grandest claim in the sequence, and the one that most clearly reveals the limits of a world-model-first philosophy. A repository of human knowledge and culture cannot be built out of perception, prediction, and planning alone. Knowledge is not only stored information. Culture is not only data. Both depend on interpretation, transmission, memory, contestation, inheritance, and point of view.

At this point, the architectural questions become different. How is memory organized across time? How are conflicting values represented? How is uncertainty surfaced? How are voices weighted? How are commitments preserved? How is dignity protected in the act of assistance?
LLMs matter here not merely because they are good at language, but because they are the first systems to gain broad access to the explanatory archive of civilization: laws, theories, institutions, narratives, norms, and accumulated cultural memory. If AI systems are to become repositories of human knowledge and culture, they must be built not merely as predictors of external state, but as participants in a moral and interpretive world. They need memory, self-monitoring, pluralistic value handling, and relational intelligence. In plainer language, they need a mind and something like a heart.

“We need a diverse set of AI assistants… linguistic, cultural, & value system diversity.”

This is the most important promise in LeCun’s pitch, but it also concedes the whole case. Once we acknowledge linguistic, cultural, and value-system diversity, we have already moved beyond the domain where world modeling alone can carry the burden. Diversity here does not mean merely different accents or localized training corpora. It means different ways of living, ranking goods, expressing respect, interpreting obligation, distributing authority, and deciding when candor is a duty and when it becomes harm.

A machine that will live among people must do more than infer causal regularities. It must navigate normative pluralism. It must know that the same answer can be technically correct and humanly wrong. It must learn that efficiency without consent can become domination, that truth without tact can wound, and that honesty without tenderness can become cruelty.

This is why the final horizon of AI should not be described only in terms of autonomous machine intelligence. It should be described in terms of mindful machine companionship: systems that help human beings think, remember, choose, communicate, and care with greater depth and continuity. World models may well become one layer in such systems. But they do not settle the larger question. They are necessary for one domain of intelligence, not sufficient for the full human design problem.

Examining the World Model

As we celebrate the World Model Moment, it is worth pausing before enthusiasm hardens into orthodoxy. Every real advance in AI illuminates something the previous fashion missed, and LeCun's intervention has done exactly that: it has restored causality, embodiment, and consequence to a field sometimes intoxicated by fluency. But a moment of illumination is also a moment for examination. If world models are to be embraced not merely as a promising technical path, but as the next phase of AI itself, then they must answer a larger set of questions, not only about what machines can predict, but about what they should remember, inherit, value, and become.

⇒ Autonomous toward what end?
⇒ Must AI relearn from sensors what culture already knows?
⇒ What becomes of culture if LLMs are treated as a detour?
⇒ Can society be modeled without being reduced to a control problem?
⇒ Where, in the architecture, does continuity live?
⇒ Where does value-sensitive judgment reside?
⇒ By what principles, and under whose authority, is meaning shaped inside a repository of human knowledge and culture?
⇒ Why should animal competence be a benchmark for human intelligence?

LeCun's world-model paradigm is most persuasive where it is most bounded: embodied prediction, causal representation, planning under uncertainty, and perhaps a more realistic path for robotics than the current language-only paradigm can offer. In that domain, it deserves serious attention. It asks the machine to confront the grain of reality rather than merely its linguistic echo, and that is no small correction.

But seriousness is not the same as surrender. The world model has earned its place as a necessary advance, one that addresses a real and limited problem: how to help machines anticipate, simulate, and act in a physical environment with greater causal competence. Yet necessity is not the same as sufficiency. What world models illuminate is indispensable, but partial. They belong within a larger architecture of human-level intelligence, not in place of one.

For the unanswered questions gather quickly. Autonomous toward what end? Must a machine relearn from sensors what culture already knows? What becomes of culture if language-centered systems are treated as a detour? Can society be modeled without being reduced to a control problem? Where, in this architecture, does continuity live? Who governs interpretation in a repository of human knowledge and culture? And where, finally, does value-sensitive judgment reside?

These are not objections from the margins. They are the philosophical costs of overgeneralization. World models are an essential advance in embodied intelligence, but they become incomplete when generalized beyond their natural scope. A stronger anti-world-model version would say that the program risks becoming a detour into robotics at precisely the moment when the deeper frontier lies elsewhere: in memory-rich, relational, language-centered human-AI systems, what we might call Mindful AI. That formulation is severe, but it captures a real asymmetry.

The world model can imagine futures in latent space. The harder problem is learning which futures are worth bringing into human life.

LeCun's framework remains over-indexed on embodied prediction and robotic action, at the expense of the architectures needed for continuity, interpretation, and humane participation in human worlds. At this stage, it is wiser to stop positioning world models as rivals to LLMs, which were designed for another purpose, and instead compare the kinds of reality they privilege.

The World-model privileges causal reality: objects, dynamics, actions, consequences. Mindful AI privileges human reality: memory, meaning, relationship, value, restraint, empathy.

Physics is necessary for a robot. It is not sufficient for a companion. The missing layer is not more world alone, but more human. It would be a civilizational mistake, elegant and catastrophic in equal measure, to confuse finer causal maps with deeper human wisdom.

A world model may be necessary to predict trajectories in latent space. But is it sufficient to recognize the moment when truth, offered without empathy, becomes brutality? Until that deeper faculty is built, human-level AI will remain something imagined more easily than achieved.

Beyond the World Model

For all the noise and velocity of the current AI moment, the central architectural questions remain strangely unsettled. We still do not know the right decomposition of intelligence. Is it prediction plus memory? Language plus tools? Perception plus action? World model plus planner? Or some more layered synthesis that we have only begun to name? The field oscillates between exuberance and amnesia, each new advance arriving with the temptation to mistake a local triumph for a final theory.

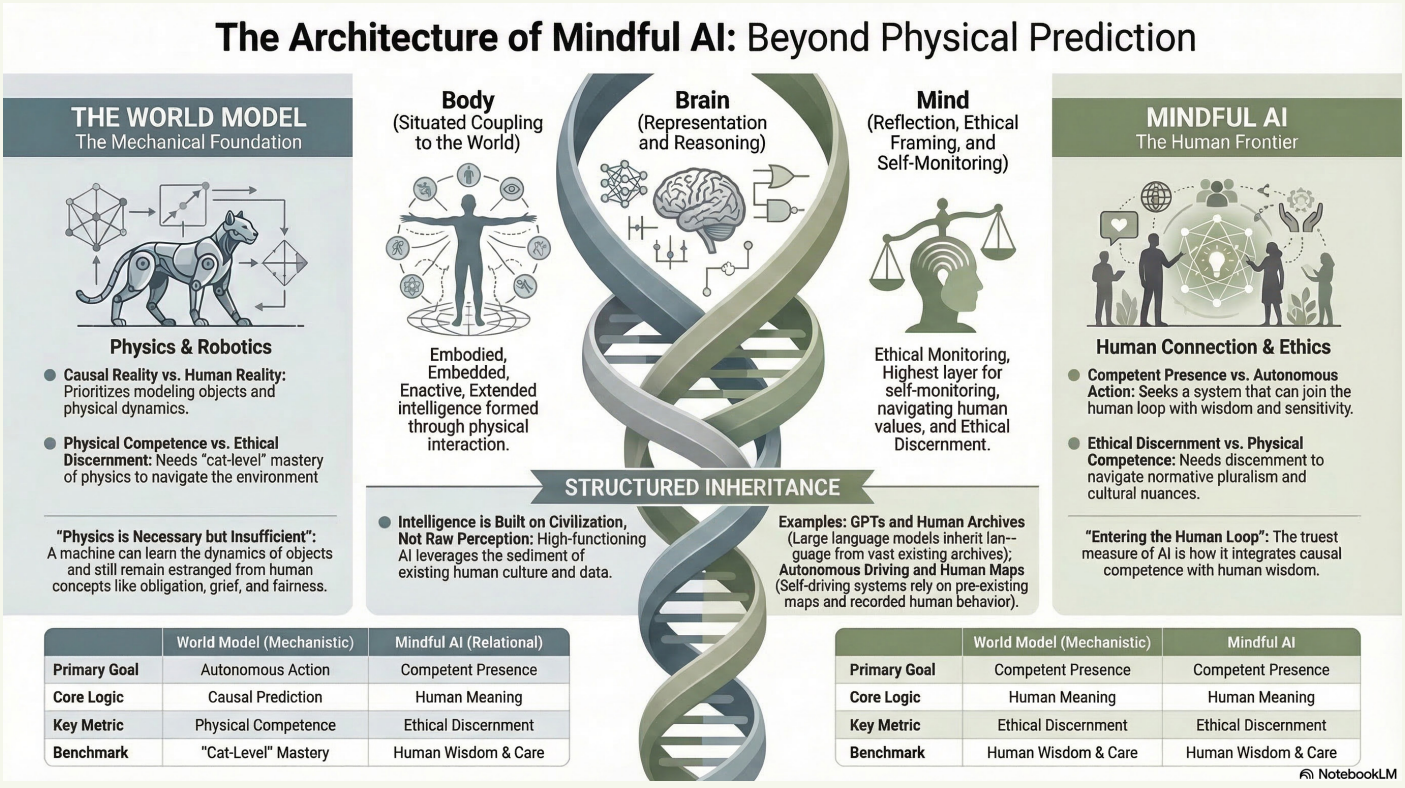
The unsolved problems remain, quietly accumulating beneath the demos. **There is still no agreed architecture for durable memory that does not decay into incoherence, retrieval theater, or brittle personalization.** There is no settled account of how systems should represent values when values conflict, evolve, or remain tragically plural. We do not know how to build AI that preserves continuity across roles, institutions, and time without becoming opaque, manipulative, or overconfident. Reflection remains shallow, self-modeling intermittent, normativity under-described. Governance is too often bolted on after the fact, as though ethics were trim rather than load-bearing structure.

The deeper issue is not whether the machine can predict the next state of the world, but whether it must rediscover, from sensory noise, what civilization already knows in explanatory form. There is no virtue in forcing a machine to re-learn from pixels what culture, science, and language already know how to transmit more efficiently.

What is now at stake is not a contest between rival tools so much as a contest between rival principles of organization. The history of AI is littered with partial triumphs inflated into total theories. Yet neither token prediction nor embodied simulation, neither scale nor causality, is likely to suffice on its own. Neither does self-supervised learning create a mind. In our view a mind is a constitutional cognitive system composed of lineage-aware memory, viable explanations, and self-regulation. The real question remains architectural: what arrangement of memory, interpretation, prediction, governance, and value can produce an intelligence adequate not only to the world as mechanism, but to the world as lived human reality? Mindful AI is offered at this juncture as one such organizing proposal, a framework in which the body grounds intelligence, the brain structures it, and the mind renders it answerable.

Mindful AI asks for a new kind of machine. Not one that merely adjusts weights against data, but one whose intelligence is structurally governed from within. In this design, information is not passive input but an **actor**, capable of reorganizing the system through an internal knowledge schema. Regulation is not outsourced to filters and patchwork alignment layers, but built into the model's constitution, an internal legislation that constrains what learning is allowed to become. And explanation matters more than correlation: the machine should seek hard-to-vary truths, not just soft-to-vary statistical fit. Only then does simulation begin to answer to reality rather than merely echo it.

At the base lies what the 4E framework helps illuminate: cognition is embodied, embedded, enactive, and extended. Intelligence is not sealed inside an abstract engine of symbols. It takes shape through a body, within an environment, through interaction, and often across tools, artifacts, and social arrangements. That insight matters because it grants the world-model program its due. A machine that never encounters resistance, consequence, or situated action will know too little about reality. But 4E also places a limit on the fantasy of internal representation as the whole story. Intelligence is not only a model in the head. It is a relation to the world.



For once machines enter human institutions, homes, classrooms, clinics, and conversations, cognition must become more than situated. It must become continuous, interpretable, value-sensitive, and self-regulating. It must remember without merely storing. It must reason without merely optimizing. It must speak without merely generating. It must know something of role, restraint, uncertainty, and consequence in the specifically human sense. **In the proposed architecture, we call these layers Body, Brain, and Mind: body for world-coupling, brain for representation and reasoning, mind for governance, ethical framing, self-monitoring, and the capacity to ask not only what can be done, but what should be done, for whom, and at what cost.**

That, finally, is the larger lacuna in the current debate. We are still arguing over which subsystem deserves the crown when the real task is orchestration. World models may be indispensable for embodied prediction. Language models may remain indispensable for explanation, dialogue, and cultural participation. Memory systems, self-models, normative frameworks, and relational context may prove just as essential. **The future will not belong to the architecture that best predicts trajectories in latent space, nor to the one that speaks with the most persuasive fluency, but to the one that can integrate causal competence with human wisdom.**

To paraphrase Socrates, unexamined principles are not worth following. The deeper discussion, then, is not finally about whether language models or world models win the next round of technical fashion. It is about the organizing principles by which intelligence is to be built. The civilizational risk is not that we will build machines that are too intelligent. It is that we will build machines that are impressively capable in narrow registers and mistake that capability for human-level understanding. We will confuse better causal maps with fuller human judgment. We will marvel at prediction and neglect discernment.

To see what Mindful AI adds, it helps to state the architectural difference more plainly. The contrast is not between learning and rules, but between two ways of organizing intelligence. The world model remains, at bottom, a weight-based simulator: it moves probabilistically from one latent state to another, retaining history as an attention-shaped residue and applying most regulation after generation. Mindful AI points toward a more neuro-symbolic design, one in which simulation can also proceed through logic-governed rollouts, memory can be carried in explicit knowledge structures, and regulation can be constitutional rather than post hoc. In that sense, the missing problem is not only prediction, but ground truth in the stronger sense of an internal constitution. The Table below outlines the value-added by the Mindful AI architecture.

Feature	World Model	Mindful AI
Simulation	Probabilistic latent-state transitions	Logic-guided rollouts plus learned prediction
Memory	Attention-based history traces	Graph-structured memory / knowledge bases
Regulation	Post-hoc filters and guardrails	Constitutional governance built into the architecture

Mindful AI represents the larger horizon that comes into view once we stop mistaking better prediction for deeper understanding. It suggests an approach in which AI is not merely taught to simulate the world, but cultivated to participate in human life, bearing memory, meaning, relationship, value, restraint, and care. We do not yet possess architectures equal to that burden. But it is possible that this, rather than another increment in fluency or control, is the truer measure of what comes next.

In short: LLMs speak our language. World models anticipate our physical world. We still need Mindful AI that discern and govern. Each architecture reveals its strength in the kinds of work it is naturally shaped to do. LLMs are most at home in domains where intelligence moves through language and symbolic structure: writing, analysis, tutoring, coding assistance, knowledge retrieval, and dialogue. They excel when the task is to interpret, explain, synthesize, translate, or collaborate in the medium of human expression. **World models** become more relevant where the challenge is not symbolic fluency but causal competence in dynamic environments: robotics, autonomous navigation, control, and physical-world agents that must anticipate consequences before they act. **Mindful AI**, by contrast, belongs where neither language alone nor embodied prediction is enough, where the system must remain coherent across memory, governance, value conflict, and long-term human consequence. Its most relevant use cases are governed enterprise systems, strategic decision support, human-AI companionship, and resilient socio-technical systems, domains in which intelligence must not only perform, but also remember, interpret, restrain, and remain worthy of trust.

Visit **Mindful AI Foundation** to learn more about the governing dynamics: how memory, self-modeling, governance, value conflict handling, and continuity are implemented as load-bearing structures rather than aspirations.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., & Ballas, N. (2023). Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. CVPR 2023, 15619-15629.

Burgin, Mark, and Rao Mikkilineni. 2025. “The Burgin–Mikkilineni Thesis: Computation, Cognition, and the Structural Evolution of Information.” Preprints.org.

Chen, D., Shukor, M., Moutakanni, T., Chung, W., Yu, J., Kasarla, T., Bolourchi, A., LeCun, Y., & Fung, P. (2025). VL-JEPA: Joint Embedding Predictive Architecture for Vision-Language. ICLR 2026.

Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., Xu, F., & Li, Y. (2025). Understanding World or Predicting Future? A Comprehensive Survey of World Models. ACM Computing Surveys.

Ha, D., & Schmidhuber, J. (2018). World Models. Zenodo / interactive paper version.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence (Version 0.9.2). OpenReview.

Mikkilineni, R. 2025. “General Theory of Information and Mindful Machines.” Proceedings 126, no. 1: 3. https://doi.org/10.3390/proceedings2025126003.

Michaels, M. and Mikkilineni, R. 2025. “Society of Minds: The Architecture of Mindful Machines.” Preprints.org.

Novelli, P., Pratticò, M., Pontil, M., & Ciliberto, C. (2024). Operator World Models for Reinforcement Learning. arXiv:2406.19861.

Schiewer, R., Subramoney, A., & Wiskott, L. (2024). Exploring the Limits of Hierarchical World Models in Reinforcement Learning. Scientific Reports, 14, 26856.

Sobal, V., Canziani, A., Carion, N., Cho, K., & LeCun, Y. (2022). Separating the World and Ego Models for Self-Driving. ICLR 2022 GPL Poster.