MINDFUL AI
FOUNDATION

The Limits of the World Model

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

Why Human-Level AI Needs More Than Physics

We are at an inflection point in AI. For the past few years, the field has been dazzled by the rise of large language models: systems that can write, reason, translate, summarize, tutor, code, and converse with uncanny fluency. In response, a counterforce has gathered momentum. Its most eloquent promoter is Yann LeCun, who argues that if we are serious about human-level intelligence, we must move beyond language and toward systems that learn how the world works through perception, action, and prediction. This is an important debate, and it matters now because the field is deciding what problem it thinks it is solving.

Is AI primarily about building autonomous agents that can act competently in the physical world? Or is it about building systems that can participate wisely in the human world: in memory, language, care, judgment, culture, and trust?

LeCun's "world model" program offers a powerful corrective to the excesses of LLM triumphalism. It reminds us, rightly, that text is not the whole world, that intelligence is not exhausted by next-token prediction, and that action requires more than eloquence. But the world model is now being overpromoted: from a solution to embodied prediction into a theory of human-level intelligence. *That is a tall claim.*

What follows, then, is not a critique of the need for world models. It is an objection to their coronation. The deeper frontier in AI is not merely the construction of better predictive machinery for the physical world. It is the cultivation of 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 possible futures, but of discerning which futures deserve a place in the world we share. The deeper frontier is not merely the construction of better predictive machinery for the external world. It is the cultivation of 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 possible futures, but of discerning which futures deserve a place in the world we share. We may as well call them ***Mindful AI***.

1. “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. That mythology deserves puncturing. LLMs are not complete minds. They do not natively possess persistent memory, robust self-modeling, grounded agency, 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 problem is not that world models are false and LLMs true. It is that they have come to represent different hungers of the field. World models seek contact with the grain of reality, with objects, consequences, and action. LLMs, imperfect as they are, have revealed another province of intelligence: the ability to move through language, inherit culture, sustain memory, and participate in the long human traffic of interpretation and exchange. Human-level intelligence is not only a matter of predicting the world. It is also a matter of entering the symbolic and cultural worlds that people have made together.

What LLMs revealed, despite all their limitations, is that the frontier of AI is not just mechanistic competence. It is relational participation. They showed that systems can already enter the space where humans ask for explanation, reassurance, reflection, help, and continuity. The failure of these systems is not that they are too linguistic. It is that they are not yet mindful enough.

2. “A path towards autonomous machine intelligence.”

That phrase contains, quietly, the first boundary line. ***The operative word is autonomous.***

Autonomy is a real engineering achievement. A robot, a vehicle, or a planning system must often 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.

That requires a different center of gravity. Not just action without supervision, but judgment with sensitivity. Not just successful state transition, but continuity of relationship. Not just competence in the world, but wisdom in the worlds that people build together.

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.

3. “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, adapt to the environment, 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, rewrite a legal clause in plain language, or remember the emotional stakes of a family decision. These are not side functions of civilization. They are central ones.

So 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. It captures one layer of intelligence, but not the full human design space. Physics is necessary for a robot. It is not sufficient for a companion.

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

This is the conceptual core of the program. ***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.

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. It may know what follows from an action without knowing what should restrain one.

The problem before us is not merely richer sensory input. It is richer human framing.

5. “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.

The phrase “any complex system” has a certain splendor to it. *It suggests an architecture broad enough to scale from perception to causality to action, perhaps even from particles to organisms to societies.* It promises not just a robot that can grasp objects, but a general engine for understanding the dynamics of the world.

Yet this is precisely where caution is needed. ***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. These are not simply hidden state variables in a universal action-conditioned model. They belong to another register.

The danger here is not technical ambition. The danger is ontological flattening. When the framework stretches from physical causality toward human systems, it risks if what matters most about a society is 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. Not everything meaningful is best represented as a control problem.

This is also where the architectural gaps become harder to ignore. 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. They may learn how a system changes. They do not yet tell us how a system should remain answerable over time.

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

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

Here the argument turns, perhaps more than it admits. 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. This is the hinge of the debate.

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.

The assistant form pulls us away from a purely robotic conception of intelligence and toward a dialogic one. The future is not simply a machine acting on the world. It is a machine joining the human loop of meaning.

7. “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. A repository of human knowledge and culture must not only contain, but also discriminate, contextualize, preserve plurality, and navigate disagreement without erasing it.

At this point, the core 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?

These are not polish layers. They are the architecture. The phrase Mindful Machines is useful here not as decoration, but as a design requirement. 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.

8. "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 compromises 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.

Both GPTs and autonomous driving began, in a fundamental sense, not from raw innocence but from inheritance. GPTs did not wake into language the way a child wakes into the world; they were trained on vast existing archives of books, websites, code, and conversation, the sediment of culture already written down. Autonomous driving followed a similar logic. It did not begin by rediscovering roads from first principles, but by drawing on existing 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. That is the deeper analogy. Intelligence, artificial or human, rarely begins in a vacuum. It begins in a world already structured by memory. That raises fundamental questions on the underpinnings of the World Model.

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

Conclusions

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. That is where admiration must mature into proportion: not a rejection of the world model, but a refusal to mistake one essential layer for the whole edifice.

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? What governs interpretation when an AI system becomes 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. The sharper anti-world-model critique 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 Machines. 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 program privileges causal reality: objects, dynamics, actions, consequences. And Mindful AI privileges human reality: memory, meaning, relationship, value, restraint, care.

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 know when honesty without empathy becomes cruelty? Until then, human-level AI will elude humanity, not because the machine knows too little physics, but because it knows too little of being human.

Epilogue: 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 what the right decomposition of intelligence is. 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 usual temptation to mistake a local triumph for a final theory. One year, scale is destiny. The next, embodiment is. Then comes agency, or memory, or multimodality, or synthetic data, each proposed not merely as an ingredient but as the long-awaited key to the whole edifice.

Yet 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. Even the more persuasive correctives to LLM triumphalism, world models, embodiment, agency, often answer one deficit by enlarging another. They restore contact with causal reality, but they do not yet explain how a machine should live among meanings.

The deeper issue is not only whether the machine can predict the next state of the world, but whether it must rediscover, from sensory noise alone, what human explanation already knows. A civilization does not begin each generation by relearning gravity from scratch. It transmits laws, meanings, and models through culture. An AI architecture that ignores this inheritance may become impressively grounded and curiously wasteful.

That is why the deeper challenge is not merely to give AI a better map of the world. It is to give it a more adequate place in the human world. This is where the idea of **Mindful AI** becomes useful, as architecture. Mindful AI would not be defined by one faculty elevated as supreme. Not by language alone. Not by world modeling alone. Not by agency, memory, or embodiment alone. It would be built as a layered system adequate to the layered nature of intelligence itself.

At the base lies what our 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. And still, even that is not enough.

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 architecture I have in mind, one might call these layers **Body, Brain, and Mind**: body for situated coupling to the world, brain for representation and reasoning, mind for reflection, 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 of AI will not be settled by prediction alone, but by whether prediction is placed in service of memory, purpose, and humane coherence. 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.

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 understanding. We will confuse better causal maps with fuller human judgment. We will marvel at prediction and neglect discernment.

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.

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.

- 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.

- 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*.