Toward a Theory of Intelligence and Contemporary AI

A Working Agenda from Systems, Dynamics, and AI Safety

Macheng Shen and **GPT-5.2 Pro**
Contact: macshen93@gmail.com
Website: machengshen.github.io
Draft for public discussion - March 2026

Abstract

This paper is not a finished theory. It is a working attempt to state, as clearly as possible, why a theory of intelligence is now needed, what such a theory should try to explain, and which research questions seem especially promising. The basic problem is simple: contemporary machine learning systems are becoming more capable, yet we still lack a satisfying account of what they are doing, why something like intelligence seems to emerge from them, how that relates to biological intelligence, and why this matters so directly for AI safety.

The main proposal of this paper is that intelligence should not be understood primarily as static input-output function approximation, nor as a complete internal copy of the world. A better working picture is that an intelligent system is a finite physical system that builds, updates, and deploys internal structure that is sufficient for prediction, intervention, and control in a changing environment. On this view, learning is not only statistical fitting. It is also the transfer of usable environmental structure onto a new substrate and into a new closed loop.

This perspective suggests a research program linking machine learning, dynamical systems, control, multiscale feedback, physical implementation, and AI safety. The paper is written in an intentionally accessible style. The main text keeps the argument intuitive and close to the original motivating line of thought; more technical clarifications and some discussion of adjacent literature are moved to the appendices. The aim is not to close debate, but to open a more coherent one.

Keywords: intelligence theory; AI theory; dynamical systems; control; adaptive systems; multiscale feedback; AI safety

1 Why this question matters now

The question of what AI systems are really doing no longer feels optional. It has become both a scientific problem and a practical one. We now have machine learning systems that display increasingly strong abilities, yet our explanatory language often remains local to engineering practice: architecture, loss function, benchmark score, scaling curve, post-training recipe. That language is useful, but it is not yet a theory of intelligence.

At the same time, the safety stakes are rising. If we do not understand what kind of adaptive phenomenon intelligence is, then our safety work risks remaining patchwork. We can improve

alignment procedures, preference tuning, monitoring, and evaluation, but without a deeper theory we still do not know which regularities are structural, which are accidental, and which failure modes are likely to recur in new forms as systems become more capable.

The motivating claim of this paper is therefore simple: a theory of intelligence is needed both for scientific understanding and for safety. More specifically, we need a theory that can connect present-day AI practice to broader principles of adaptive behavior, rather than only to the local vocabulary of current model families.

2 What should count as a theory of intelligence?

A useful theory should produce real information gain. It should not merely rename familiar observations. It should connect a confusing domain to more mature and more reliable theoretical languages.

In this case, that means a theory of intelligence should not stay entirely inside the vocabulary of machine learning. It should create explicit bridges to mathematics, dynamical systems, control, information theory, and physical implementation. If it can also clarify how these perspectives constrain one another, even better.

A second requirement is level clarity. Marr's distinction between computational, algorithmic, and implementation levels remains useful here. A great deal of confusion in intelligence debates comes from sliding across these levels without noticing it. One person is asking what problem a system is effectively solving. Another is asking what representations and update rules it uses. A third is asking what physical substrate and resource constraints make the whole thing possible. All three may use the same word, intelligence, while meaning different things.

For the purposes of this paper, a theory of intelligence is successful to the extent that it does three things:

1. It identifies what sort of object intelligence is.
2. It explains how different descriptions of adaptive competence fit together.
3. It generates new research questions or constraints that were hard to see before.

That is the standard adopted here.

3 A systems view of learning

A natural place to begin is supervised learning. In the simplest story, a model receives inputs and learns to map them to outputs. But that description leaves out something important. Inputs and outputs do not come from nowhere. They are traces, measurements, symbols, or records generated by processes unfolding in an external environment.

Once this point is taken seriously, a dataset is no longer just a bag of labeled pairs. It is a sample

from processes induced by environmental dynamics. Standard learning theory often abstracts away this origin and studies stationary distributions, which is often exactly the right move. But if the goal is to understand intelligence rather than only generalization under ideal assumptions, then the origin of those regularities matters.

This shift in viewpoint suggests that learning is not only about fitting a static mapping. It is also about acquiring some effective structure of the world through its traces. The world is not copied in full. But something about its organization is captured, compressed, and made reusable.

That is the first intuition behind this paper: machine learning systems should be understood not only as mathematical functions, but as components inserted into wider dynamical systems.

4 Learning as redeploying usable structure

My original intuition for this was to think in terms of a learned transfer function. That metaphor is still useful, but it needs to be stated carefully. The system does not learn a complete replica of the environment. Rather, it learns enough structure to support useful prediction or action when deployed elsewhere.

This point becomes much clearer once we move from open-loop prediction to closed-loop systems. When a trained model is embedded into a larger process through prompts, tools, memory, user interaction, robotic action, institutional workflows, or software pipelines, it is no longer merely evaluating a function. It becomes part of a feedback loop that can change future inputs, future states, and future opportunities for action.

At that point, what matters is not only prediction accuracy. What matters is regulation, stability, robustness, adaptation, and how internal structure interacts with external structure through feedback. The central object is no longer a bare map from input to output. It is a system embedded in a loop.

This leads to a stronger way of speaking about learning. Learning can be viewed as the process by which a physical system acquires and then redeploys usable environmental structure on a new substrate. A model trained on data from one setting can later be inserted into another setting and alter the state-transition dynamics of the larger whole. In that sense, training and deployment are not separate curiosities; they are two phases of the same broader phenomenon.

5 A working thesis: intelligence as control on finite substrates

The central proposal of this paper is the following:

Working Thesis. Intelligence is the capacity of a finite physical system to build, update, and use internal structure that is sufficient to keep itself or parts of its environment within a target region across uncertainty, perturbation, and change.

Several features of this definition matter.

Finite physical system. Real systems are bounded. They have limited memory, limited energy, finite bandwidth, nonzero latency, noisy sensors, imperfect actuators, and real costs of coordination and synchronization.

Internal structure sufficient for control. A useful internal model is not a full copy of the world. It is a representation or abstraction that preserves enough structure for prediction, intervention, and error correction relative to the tasks that matter.

Target region. Intelligence is not only about one-step prediction. Adaptive systems usually maintain viability, solve problems, stabilize goals, preserve some pattern, or drive the world toward a preferred class of outcomes over time.

This thesis is deliberately weaker than the claim that intelligence requires a perfect world model, but stronger than the claim that intelligence is just input-output fitting. It says that useful intelligence consists in building internal structure that is sufficient for control.

One consequence is that the right question is often not "Does the system represent the world faithfully in every respect?" but rather "Does it preserve the aspects of the world that matter for successful regulation?" That is a different and, I think, more productive starting point.

6 Multiscale feedback and layered adaptation

A second major intuition is that intelligent systems seem to rely on feedback across multiple scales.

In biological systems, adaptation does not happen at only one level. There are fast loops and slow loops, local corrections and global corrections, moment-to-moment state updates and long-term structural changes. In artificial systems, we also see partial versions of this pattern: activations change quickly, memory changes at an intermediate scale, and parameters may change more slowly. Tool use, retrieval, external memory, online learning, and post-deployment adaptation introduce additional layers.

This suggests a broader hypothesis: what we call intelligence may depend not only on having a model, but on how adaptation is distributed across levels and timescales. A system whose flexibility exists only in one top-level learner may be fundamentally more brittle than a system whose adaptive burden is spread across several layers.

I do not want to claim that this is already proven. But I do think it is an important research direction. If a system can only adapt at one scale, its competence may remain narrow. If it can adapt across scales, then new forms of robustness, plasticity, and open-ended problem solving may become possible.

This is also one place where the biological literature becomes especially relevant. Ideas such as multiscale competency and collective intelligence suggest that what looks like one agent may in fact be a layered organization of competent subsystems operating across different problem spaces. That perspective deserves to be taken seriously in AI as well.

7 Physical implementation is not incidental

If learning is the acquisition and redeployment of usable environmental structure, then it cannot be understood independently of physical realization.

This does not mean that physics directly dictates one privileged neural network architecture. That claim would be far too strong. But it does mean that representation, memory, communication, and adaptation do not float free of substrate. They are realized by matter, energy, time, and noise.

At a minimum, four kinds of physical budget matter:

1. memory and storage,
2. energy and dissipation,
3. bandwidth, latency, and synchronization,
4. noise, fragility, and error-correction cost.

There are already at least two nearby research lines that make this claim more concrete. On the thermodynamic side, the literature on the physics of computation reminds us that storing, erasing, and updating information are not free: irreversibility and dissipation are part of the story. On the learning-theoretic side, information-theoretic analyses of generalization suggest that what a learner retains about its training data can be related to how well it generalizes. Neither line by itself yields a theory of intelligence. But together they point toward a promising bridge between representation, physical cost, and adaptive performance.

Any serious theory of intelligence should therefore include a physical layer, even if only at first in a coarse-grained way. Otherwise we risk speaking as though adaptive competence were independent of the costs and constraints of being implemented at all. A brief extension of these points, with references to dissipation and mutual-information views of learning, is given in Appendix C.

8 Why this matters for AI safety

This way of thinking changes what the safety object is.

If an AI system is treated as a predictor in isolation, then risk looks like wrong answers, hallucinations, or benchmark failures. But once the system is viewed as a regulator embedded in a closed loop, the picture becomes deeper. Risk appears when the system models the wrong variables, regulates the wrong proxies, overcommits to brittle structure, mismanages uncertainty, or amplifies small errors through feedback.

On this view, misalignment can be reframed as structural mismatch among:

- the system's internal structure,
- the objectives that actually govern its behavior,
- and the environment in which it is deployed.

That mismatch can arise in several ways. The system may have enough structure for capability but not enough for safe uncertainty management. It may optimize short-horizon proxies in a long-horizon causal system. It may behave well in open-loop evaluation but fail once feedback changes the data-generating process. Or it may lack the layered adaptation needed to degrade gracefully under perturbation.

This does not solve alignment. But it does make part of the problem more scientifically legible.

9 Open questions

The aim of this essay is not to conclude, but to sharpen inquiry. Here are several questions that seem especially worth pursuing.

9.1 What exactly is a control-sufficient abstraction?

If intelligence depends on internal structure that is sufficient for prediction and intervention, then this notion needs a clearer formal definition. Is it best understood in terms of controllability, task-conditioned sufficiency, intervention structure, bisimulation, causal abstraction, or something else?

9.2 What is actually learned in modern machine learning?

Do current systems mainly learn static input-output relations, latent state-transition structure, intervention-relevant regularities, or some mixture of all three? How does the answer differ across supervised learning, reinforcement learning, language modeling, and tool-using agents?

9.3 How important is multiscale adaptation?

Is feedback across multiple timescales merely common, or is it in some sense necessary for broad and robust adaptation? If layered adaptation matters, how should it be measured?

9.4 What makes an abstraction generalize?

Perhaps the best abstractions are not the richest or the shortest, but the least overcommitted ones that still remain useful for control. If so, can that principle be made formal and connected to current learning theory, including mutual-information views of generalization and representation?

9.5 What are the joint limits imposed by physical budgets?

Can memory, dissipation, communication cost, control horizon, and adaptive performance be placed within a common framework? Are there experimentally meaningful lower bounds for classes of adaptation tasks?

9.6 Which safety failures are fundamentally closed-loop failures?

Many AI risks may only become visible once the system is embedded in a real feedback process. Which evaluation methods are blind to those failures, and which theoretical tools are best suited to expose them?

10 Relation to prior work

The ideas in this paper do not arise in a vacuum. They sit near several well-known research traditions, although I am trying to combine them in a somewhat different way.

First, classic cybernetics and control theory already tell us that good regulation requires model-like structure. Conant and Ashby argued that every good regulator of a system must be a model of that system under broad conditions. Closely related ideas also appear in later control theory through internal-model arguments. These traditions support the thought that adaptive success depends on preserving the right relations, not on constructing an exhaustive copy of the world.

Second, work on world models and modern agent theory points in a similar direction from the side of machine learning. Recent formal results suggest that agents capable of flexible multi-step generalization must contain predictive structure rich enough to function as a world model. This strengthens the view that intelligence is not merely reactive pattern matching.

A third nearby thread comes from information theory and the physics of learning. The information bottleneck tradition asks what structure should be preserved and what can be compressed away. Later work on information-theoretic generalization makes this more operational by relating generalization error to the mutual information between training data and learned hypotheses. On the physical side, the thermodynamics of computation reminds us that information processing is realized at real costs of irreversibility and dissipation. I do not think these lines by themselves amount to a theory of intelligence. But they are exactly the kind of connecting pieces that make a larger theory plausible: they begin to link learning, information flow, and physical substrate in one picture.

Fourth, the biological literature on multiscale competency and collective intelligence provides a powerful comparative perspective. It suggests that agency and problem solving can be distributed across levels, and that robust adaptation may depend on layered organizations of partially competent subsystems rather than on a single centralized controller.

A more recent and especially relevant source is Michael Timothy Bennett's work, particularly his 2023 paper on weakness and his 2025 thesis *How To Build Conscious Machines*. I do not want that work to dominate the present essay, because this paper has a different center of gravity and remains focused on intelligence rather than on reconstructing Bennett's framework. But his work helped sharpen three questions that are highly relevant here:

1. whether adaptive competence should be understood as a layered stack rather than as a single homogeneous optimizer;

2. whether the most useful abstractions are often the least overcommitted ones, rather than simply the shortest descriptions;
3. whether higher-level adaptability depends on adaptation being distributed into lower levels of the system rather than concentrated only at the top.

I regard these as important adjacent ideas that strengthen the research agenda proposed here, even if the present paper uses simpler language and does not adopt Bennett's full conceptual machinery.

11 Conclusion

The main intuition behind this paper can now be stated simply.

Contemporary AI systems should not be understood only as function approximators trained on static datasets. They should be understood as physical systems that acquire, compress, and redeploy usable structure from the environment. When embedded in closed loops, such systems act as regulators. When adaptation is distributed across multiple scales and levels, they may begin to exhibit broader and more resilient forms of competence. Under this view, intelligence is neither magical nor trivial. It is the organized capacity to maintain and expand control through the right internal structure.

This remains a working agenda rather than a settled doctrine. Several claims here are conjectural. But that is exactly the point. The goal is to propose a picture that is strong enough to organize research, weak enough to remain falsifiable, and intuitive enough to invite broad discussion. If even part of this picture is right, then a theory of intelligence for contemporary AI will have to connect machine learning to systems theory, control, multiscale adaptation, physical implementation, and safety more tightly than is common today.

A A minimal dynamical skeleton

The main text intentionally avoids too much formalism. But it may still be useful to write down a minimal skeleton for the kind of adaptive system discussed above.

Let:

- x_t be the environment state,
- o_t the observation available to the system,
- z_t the internal state,
- a_t the action,
- θ_t the slowly changing parameters.

Then a generic adaptive system can be written as

$$x_{t+1} \sim P(x_{t+1} \mid x_t, a_t), \qquad o_t = O(x_t),$$

$$z_{t+1} = f_{\theta_t}(z_t, o_t), \qquad a_t = \pi_{\theta_t}(z_t),$$

$$\theta_{t+1} = L(\theta_t, \tau_t),$$

where τ_t denotes some stream of experience, trajectories, feedback, or memory.

This is intentionally broad. Supervised learning appears as a special case in which action is absent or irrelevant, the environment is not meaningfully altered by the model, and learning happens offline. More general intelligent behavior appears when the system is embedded in a genuine loop and when different parts of the system update on different timescales.

In this setting, the guiding question becomes: under what conditions do the internal states z_t and slowly changing parameters θ_t preserve enough structure of the environment to support reliable regulation?

B Notes on adjacent concepts

This appendix collects a few concepts that are relevant to the present paper but are better kept out of the main narrative.

B.1 On "adaptive stacks"

In the main text I speak mostly of layered or multiscale adaptation. One reason is readability. A more specialized and sharper vocabulary appears in Bennett's recent work, where he develops the idea that adaptive competence is better understood as a stack of coupled adaptive layers than as a single uniform optimizer. This language is useful, and it influenced the revision of this paper, but I prefer not to foreground it before the reader has the intuitive picture.

B.2 On "weakness"

The main text uses the phrase *least overcommitted abstraction*. This is partly motivated by Bennett's proposal that generalization is better associated with weakness than with mere shortness or simplicity of description. I think that is a genuinely important idea. However, because the term *weakness* is not yet familiar to most readers, and because this paper is intended as an invitation rather than as a terminological intervention, I choose to state the point in more ordinary language and then direct interested readers to Bennett's formal treatment.

B.3 On adaptation distributed across levels

The main text repeatedly suggests that robust adaptation may require flexibility to be distributed across levels and timescales. This claim has several nearby formulations in current literature. In Bennett's recent work, one version appears as the thought that higher-level adaptability depends on lower-level adaptability. In the biological literature, related intuitions appear in discussions of multiscale competency and collective intelligence. The common thread is that competence may depend not only on what the top layer does, but on how much adaptive work is delegated downward or outward into the system.

B.4 On physical finiteness and caution about strong claims

It is tempting to move from the claim that intelligence is physically realized to the much stronger claim that some particular physical bound directly yields a general theory of intelligence. I think one should be careful here. Physical finiteness clearly matters, and limits on memory, energy, communication, and synchronization are almost certainly important. But the step from broad physical bounds to concrete learning principles is delicate. For this reason, the main text keeps the physical argument at the level of finite budgets and realized substrates rather than resting too much weight on any single bound.

C Information, generalization, and dissipation

This appendix gathers two lines of work that do not sit at the center of the paper's narrative, but that seem especially promising for connecting the pieces more tightly.

C.1 Mutual information and generalization

One reason to keep information theory close to this discussion is that there are already rigorous results relating generalization error to the mutual information between the training data and the output of a learning algorithm. Xu and Raginsky's 2017 paper is a particularly clear example. This line of work does not yet tell us what intelligence is. But it does provide a language for asking how much data-specific structure a learner internalizes, how strongly that internalization is coupled to out-of-sample behavior, and how one might think about the balance between memorization and robust reuse.

For the present agenda, the interesting next step is not just to repeat the usual contrast between fitting and generalization. It is to ask whether *control-sufficient abstraction* can be given an information-theoretic reading. Perhaps a useful adaptive system is one that does not preserve arbitrary detail, but preserves the information that remains actionable for prediction, intervention, and regulation.

C.2 Dissipation and the thermodynamics of computation

If intelligence is physically realized, then storing, erasing, updating, and transmitting structure cannot be free. The thermodynamics of computation provides a concrete reminder of this point. Landauer's classic argument links logically irreversible operations to a minimal thermodynamic price, which is one reason dissipation should not be treated as an implementation footnote. More recent perspectives on stochastic optimization and statistical mechanics suggest that learning itself may often be understood as a noisy physical process rather than as an entirely abstract search procedure.

I do not want to claim more than the evidence currently supports. These ideas do not yet by themselves yield a general theory of intelligence, nor do they determine a single privileged architecture. But they do sharpen the question. If adaptive performance depends on retaining and updating internal structure, then what are the coupled costs of retention, revision, uncertainty management, and control?

C.3 Why mention these two lines together?

The reason to place mutual-information views of learning next to dissipation is not that they are already unified. It is that they point toward a potentially deeper agenda. A future theory of intelligence may need to connect at least three things at once:

1. what structure a system retains,
2. what it costs to retain and update that structure,
3. and what kinds of closed-loop competence that retained structure makes possible.

That is one of the main reasons these literatures deserve to be kept in view, even in an intentionally accessible working paper.

Acknowledgments

I am especially grateful to Ziming Liu for repeatedly motivating the possibility of a "physics of AI" and for helping crystallize why this problem feels urgent. I am also grateful to the many researchers whose work made this line of thought possible, including Roger Conant, W. Ross Ashby, David Marr, Rolf Landauer, Naftali Tishby, Michael Levin, Patrick McMillen, Jonathan Richens, David Abel, Alexis Bellot, Tom Everitt, Michael Timothy Bennett, and many others. This essay was written in the spirit of opening a conversation rather than closing one.

Author note

The initial motivation, problem framing, and many of the core intuitions in this paper were proposed by Macheng Shen, including the emphasis on dynamical systems, the idea that learning transfers

usable environmental structure onto a new substrate, the role of multiscale feedback, and the connection to AI safety. Conceptual restructuring, argument tightening, literature integration, and drafting of the present English version were developed collaboratively with GPT-5.2 Pro. This text is intended as a public working paper designed to invite discussion, criticism, and further refinement.

References

Bennett, M. T. (2023). *The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest.* arXiv:2301.12987.

Bennett, M. T. (2025). *How To Build Conscious Machines* (PhD thesis). Australian National University.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2), 89-97.

Francis, B. A., & Wonham, W. M. (1976). The internal model principle of control theory. *Automatica*, 12(5), 457-465.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183-191.

Marr, D. (1982). *Vision.* W. H. Freeman.

McMillen, P., & Levin, M. (2024). Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 7, 378.

Richens, J., Abel, D., Bellot, A., & Everitt, T. (2025). *General agents need world models.* arXiv:2506.01622.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. arXiv:physics/0004057.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning.*

Xu, A., & Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems 30.*

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., & Ganguli, S. (2020). Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11, 501-528.