Toward a Theory of Intelligence and Contemporary AI

A Working Agenda for a Physics of AI

Macheng Shen *and GPT-5.2 Pro*
Equal contribution in the preparation of this working paper.
Contact: macheng93@gmail.com
Website: machengshen.github.io
March 2026

Abstract

This essay proposes a working research agenda rather than a finished theory. Its starting point is simple: contemporary machine learning systems display increasingly strong and sometimes surprising capabilities, yet we still lack a satisfactory framework for understanding what these systems are doing, how their capabilities relate to broader notions of animal and human intelligence, and why this matters so directly for AI safety. The central thesis developed here is that intelligence is not best understood as a complete copy of the world, nor merely as input-output function approximation. A more useful working picture is that an intelligent system can be viewed as a resource-bounded physical system that builds, updates, and deploys a control-sufficient compressed model of environmental dynamics inside a closed loop. On this view, learning is not only statistical fitting; it is also the transfer of usable environmental structure onto a new physical substrate. This perspective motivates a set of linked research questions at the intersection of statistical learning, control theory, information bottlenecks, thermodynamics of computation, multiscale biological organization, scaling laws, and AI safety. The aim of the paper is to articulate and justify those questions clearly enough that they can be attacked, refined, or refuted.

Keywords: *intelligence theory; machine learning; dynamical systems; control; information bottleneck; multiscale feedback; AI safety*

1. Introduction: why a theory is needed now

The recent trajectory of AI has turned what once looked like a largely philosophical question into a practical scientific one. What, exactly, are current machine learning systems doing when they appear to reason, plan, adapt, or generalize? In what sense are these abilities continuous with, or discontinuous from, the broader phenomenon we call intelligence in animals and humans? And why does answering these questions matter so much for safety, alignment, and governance?

My motivating intuition is that we are currently in an uncomfortable position. On the one hand, the capabilities of machine learning systems keep improving. On the other hand, our explanatory language is often too thin: we can describe architectures, objectives, and benchmark performance, but we do not yet have a satisfying theory of intelligence that connects present-day AI practice to more general principles of adaptive behavior. Without such a theory, our picture of safety remains incomplete. We can add patches, guardrails, preference-tuning procedures, and monitoring tools, but we still do not understand the deeper regularities governing why adaptive competence appears, how it scales, and under what conditions it becomes stable, brittle, or dangerous.

This essay is therefore organized around a simple claim: a useful theory of intelligence should connect contemporary AI to mature scientific languages rather than remain enclosed within the local

vocabulary of machine learning. In particular, intelligence should be brought into explicit relation with dynamical systems, control, information, and physical implementation. The goal is not to solve consciousness, phenomenology, or the philosophy of mind. The scope here is narrower and more operational: to understand intelligence as a phenomenon of adaptive regulation in resource-bounded physical systems.

2. What would count as a theory of intelligence?

A good theory should generate information gain. By that I do not mean merely relabeling familiar phenomena. I mean that a theory should discover structurally important relations between what previously looked like separate domains. In that sense, a theory of intelligence should not remain an internal commentary on AI engineering. It should connect intelligence to mathematics, physics, systems theory, information theory, and control.

There is also a useful methodological lesson in Marr's three levels of analysis: the computational level asks what problem is being solved, the algorithmic level asks how it is represented and computed, and the implementation level asks how the process is physically realized (Marr, 1982). A theory of intelligence that remains on only one of these levels is unlikely to be adequate. Many debates become confused because they collapse these levels. One person speaks about function, another about mechanism, and a third about hardware constraints, while all use the same word "intelligence."

For the purposes of this paper, I take a theory of intelligence to be successful to the extent that it does three things at once. First, it clarifies what sort of object intelligence is: not a mystical essence, but a class of adaptive processes. Second, it explains how different existing descriptions relate to one another: for example, how statistical learning, world modeling, planning, homeostasis, and control fit together. Third, it yields new research questions or constraints that would not have been visible without the theory. In a category-theoretic spirit, the most informative part of a theory may lie less in isolated labels than in the relations it preserves across descriptions.

3. From datasets to environment-induced processes

A natural starting point is supervised learning. At first glance, a supervised model simply maps inputs to outputs. But this description is thinner than it first appears. Inputs and outputs are not isolated symbols floating in a vacuum. They are traces, records, or projections of states generated in an external world. Their statistical structure is there because the world itself has structure.

This observation matters because it shifts the interpretation of training data. Instead of viewing a dataset merely as a bag of labeled pairs, we may view it as a sample from processes induced by environmental dynamics. Standard statistical learning theory often abstracts this situation into an independent and identically distributed or stationary setting. That abstraction is useful and often necessary, but it can also hide what is philosophically and scientifically interesting: the model is learning from structure that did not originate inside the model. It is exploiting regularities that were first instantiated in the environment.

A motivating metaphor, admittedly rough, is that training "steals" an effective transfer function from the environment and redeploys it elsewhere. The more careful claim is weaker and more defensible. A trained system does not copy the world in full. Rather, it acquires an effective internal structure

that captures enough of the world's regularities to support prediction or action in a new substrate. Once deployed, this structure no longer lives only in a theorem about generalization. It enters a new causal loop and begins to affect the world through outputs, tools, actions, and feedback.

4. A working thesis: intelligence as control-sufficient compressed modeling

The central claim of this essay is the following:

> **Working thesis.** Intelligence is the capacity of a resource-bounded physical system to build, update, and use internal states that form a control-sufficient compressed model of environmental dynamics, such that the system can keep itself or parts of its environment within a target set across perturbations, uncertainty, and distributional change.

Several terms in this sentence need emphasis.

Resource-bounded. Real systems have limited memory, finite energy, finite bandwidth, finite time, noise, and imperfect sensors and actuators.

Compressed model. The internal model is not a total duplicate of the world. It is selective. It keeps what matters for prediction, intervention, and error correction, while discarding other information.

Control-sufficient. What matters is not descriptive completeness but task relevance. The model is "good enough" insofar as it supports successful regulation and adaptation in the presence of disturbances.

Target set. Intelligence should not be defined only by next-step prediction. Adaptive systems are usually trying to keep certain variables, structures, or outcomes within viable ranges. The relevant target set need not be externally specified or consciously represented; it may consist of viability constraints, learned preferences, epistemic aims, or homeodynamic setpoints.

This thesis resonates with several existing traditions without collapsing into any one of them. Classical cybernetics suggests that successful regulation generally requires model-like internal structure. Conant and Ashby's famous result states that every good regulator of a system must be a model of that system under broad conditions (Conant & Ashby, 1970). Information bottleneck ideas sharpen the compression side of the story by distinguishing between total information and task-relevant information (Tishby, Pereira, & Bialek, 2000). Recent formal work on agency goes further and shows that agents capable of flexible multi-step goal-directed generalization must encode predictive models of their environment, even when that model is not explicitly exposed as a separate module (Richens, Everitt, & Abel, 2025).

Taken together, these lines of work suggest a more precise version of the original intuition. What matters is not whether an agent contains a complete internal copy of the world. What matters is whether it contains model-equivalent structure sufficient for successful regulation.

5. A minimal mathematical scaffold

To keep the discussion grounded, it helps to write a minimal dynamical scaffold. Let

- x_t denote the environment state,
- o_t denote the observation available to the system,
- z_t denote the system's internal state,

- a_t denote the action taken by the system,
- theta_t denote longer-timescale parameters, memory, or structural degrees of freedom.

Then a broad class of adaptive systems can be described by four coupled update rules:

1. Environment dynamics: x_(t+1) ~ P(x_(t+1) | x_t, a_t)

2. Internal state update: z_(t+1) = f_theta(z_t, o_t)

3. Policy or control law: a_t = pi_theta(z_t)

4. Learning or adaptation rule: theta_(t+1) = L(theta_t, trajectory_(0:t))

This is not yet a theory of intelligence, but it does clarify the object of study. Supervised learning appears here as a limiting case in which action does not materially affect future observations, the learning rule is mainly offline, and the causal loop is partially cut. General intelligence, by contrast, seems closer to the full closed-loop case.

This scaffold also helps sharpen a thought that originally appears as talk of "inverse transfer functions" or identity relations. The stronger and more mathematically plausible question is not whether the agent reproduces the world's transfer function exactly. It is whether there exists a representation map phi and internal dynamics T_hat_a such that, on a task-relevant quotient of state space,

phi o T_a approximately equals T_hat_a o phi.

In words: the internal model need not mirror the environment at full resolution, but it should preserve enough structure that controlled environmental transitions and internal model transitions approximately commute. This would turn the vague intuition of "copying dynamics" into a more precise question about effective theories, coarse-graining, and representation.

6. Why deployment matters: from open-loop prediction to closed-loop regulation

One reason current discourse about AI often feels incomplete is that it over-focuses on training and under-focuses on deployment. During training, we often analyze prediction error, optimization dynamics, and generalization. During deployment, however, a model enters a new causal graph. Its outputs influence future inputs through humans, tools, institutions, robots, or markets. At that point, the object is no longer just a predictor. It is a regulator inside a closed loop.

This is where control theory becomes unavoidable. In a closed loop, the relevant questions include stability, robustness, gain, observability, controllability, delayed feedback, uncertainty propagation, and failure under perturbation. The dangers of advanced AI systems often arise not because a single prediction is wrong, but because small errors or mis-specified objectives are amplified through repeated interaction.

This does not mean that intelligence should be reduced without remainder to explicit control. Exploration, curiosity, play, imagination, and representation-building are also central. But even these activities can often be interpreted as enlarging future control, reducing epistemic vulnerability, or improving the quality of later regulation. The point of the present framework is not to flatten intelligence into a narrow engineering notion of command; it is to insist that adaptive competence is easiest to analyze when prediction, memory, action, and target-maintenance are studied together.

This is also why several seemingly different traditions begin to converge. Reinforcement learning and optimal control can be written in probabilistic inference language, highlighting the deep relation between acting and inferring (Levine, 2018). Active inference and the free energy principle similarly attempt to place perception, learning, and action inside a unified optimization story (Zhang & Xu, 2024). One does not need to endorse every ambition of these frameworks to learn from them. Their shared lesson is enough: intelligence is not merely function approximation in isolation. It is adaptive behavior in a loop.

7. Multiscale feedback as a candidate structural requirement

A recurring intuition in my original line of thought is that genuinely intelligent systems almost always seem to involve multiscale feedback. Fast variables interact with slow variables. Local correction interacts with global control. Short-term responses interact with longer-term memory, development, or structural adaptation.

This intuition appears in biology very strongly. McMillen and Levin argue that living systems exhibit multiscale competency architectures in which cells, tissues, organisms, and swarms solve problems in different state spaces while collectively producing robust adaptive behavior (McMillen & Levin, 2024). That picture is highly suggestive for AI. It implies that intelligence may not be a property of a single flat optimization process. It may instead depend on several interacting layers of adaptation operating at different time scales and levels of abstraction.

At present, I would state this carefully as a conjectural direction rather than a theorem: perhaps open-ended adaptation requires at least two or more coupled plasticity scales, together with channels by which errors or objectives at one scale reshape the dynamics at another. If that is right, then many current architecture choices are not just engineering conveniences. They are partial implementations of a deeper structural requirement.

This line of thought also helps distinguish different kinds of "memory." Parameters store slow regularities. Working memory stores transient context. External tools and databases provide long-timescale support outside the original substrate. Biological morphology may itself encode task-relevant structure. A theory of intelligence should explain how these memory forms compose.

8. Physical constraints: beyond the slogan that intelligence must obey physics

If intelligent systems are physical systems, then learning and adaptation must obey physical constraints. But this claim only becomes useful when it is made specific. Saying that learning "cannot violate physics" is too weak. Almost any computation satisfies that statement. What matters are tighter constraints tied to information processing itself: dissipation, erasure cost, time cost, bandwidth, noise, locality, and storage.

Landauer's classic result provides a canonical starting point. Logical irreversibility is associated with physical irreversibility and minimal heat generation (Landauer, 1961). This does not directly derive the architecture of intelligence, but it firmly links information handling to thermodynamic cost. The broader lesson is that learning is not an abstract movement in parameter space. It is a physically realized process by which usable structure from the environment is written into a substrate.

This perspective also makes contact with modern learning theory. Xu and Raginsky show that the mutual information between training data and learned parameters can upper-bound generalization

error in broad settings (Xu & Raginsky, 2017). That result does not say that "low information means more intelligence." But it does show that one can quantify how much information has been carried from data into a learned object, and how that quantity matters for robustness to overfitting. Meanwhile, statistical-mechanical analyses of deep learning treat optimization as a dynamical process shaped by landscape geometry, effective temperature, noise, and scale (Bahri et al., 2020).

These literatures are still far from a unified theory of intelligence. Yet they already point toward a productive research triangle: information, control, and thermodynamic cost. A credible "physics of AI" will likely need to develop that triangle rather than relying on high-level appeals to energy conservation alone.

9. Emergence, scaling, and threshold phenomena

One of the strongest empirical motivations for a theory of intelligence comes from the apparent emergence of new capabilities as models scale. But this issue needs careful treatment. There are at least two live possibilities.

The first is that some capabilities really do correspond to threshold phenomena. Internal representations improve gradually, but once they cross a task-dependent horizon in compositional depth, memory span, or control fidelity, a qualitatively new macro-behavior becomes visible.

The second is that many cases of "emergence" are artifacts of measurement. Kaplan et al. show that language model loss often follows smooth power-law scaling across model size, data, and compute (Kaplan et al., 2020). Schaeffer, Miranda, and Koyejo argue that several reported emergent abilities weaken or disappear when one changes the metric or obtains better low-scale measurements (Schaeffer, Miranda, & Koyejo, 2023).

The most defensible position, in my view, is neither magical discontinuity nor blanket debunking. A better framing is that many apparently sudden abilities may arise when smooth representational improvements are observed through thresholded, nonlinear, or sparse metrics. The scientific task is then to identify which discontinuities are genuine properties of the underlying system and which are properties of our measurement procedures.

This matters for theory because it suggests that "emergence" should be studied as a relation among representation, task, and metric. That is already a more tractable scientific program than treating emergence as a mysterious primitive.

10. Why this matters directly for AI safety

The proposed framework matters for safety because it changes what the object of concern is. If a deployed AI system is a regulator embedded in a causal loop, then risk is not primarily a matter of isolated wrong answers. Risk arises when a system models, values, or controls the wrong variables, when its uncertainty is miscalibrated, when distributional shift interacts badly with feedback, or when closed-loop amplification turns small internal errors into large real-world effects.

On this view, misalignment can be described as structural mismatch among three things: the system's internal model, the objective that actually governs its behavior, and the environment in which it is deployed. That mismatch may appear at several levels. The system may represent proxies rather than the variables humans care about. It may optimize short-horizon rewards while operating in long-

horizon causal systems. It may perform well in open-loop benchmarks while failing under feedback. Or it may have learned a world model that is good enough for capability yet badly insufficient for safe uncertainty management.

This reframing does not solve alignment, but it makes part of the problem more concrete. A safety science grounded in this perspective would ask: what variables is the system effectively regulating, what model-equivalent structure supports that regulation, on what time scales does it adapt, where does uncertainty enter, and what physical or informational budgets limit correction?

11. Conjectures and research questions

The purpose of this essay is not to present a complete theory but to articulate a set of justified conjectures and open questions. I list four here.

11.1 Control-sufficient model conjecture

Any agent that can robustly achieve a nontrivial family of goals under perturbation must encode a control-sufficient predictive model of its environment, whether explicitly or implicitly.

Research questions

- How should "control-sufficient" be defined mathematically?
- Can it be quantified using information-theoretic sufficiency, controllability, or task-conditioned bisimulation?
- Under what assumptions can a world model be extracted from a policy in a canonical way?

11.2 Multiscale feedback conjecture

Open-ended adaptation requires more than one coupled time scale of state update or plasticity, together with channels that propagate constraints or errors across scales.

Research questions

- What is the minimal multiscale architecture required for robust adaptation?
- Are there impossibility results for single-timescale agents in nonstationary or partially observed environments?
- How do parametric memory, working memory, tool use, and external memory divide labor across scales?

11.3 Emergence-threshold conjecture

Many new capabilities correspond not to the sudden creation of a new faculty, but to the crossing of task-dependent thresholds in representational fidelity, compositional depth, or control horizon.

Research questions

- Can we identify task families with measurable thresholds that are independent of arbitrary benchmark metrics?
- How can we distinguish true dynamical phase transitions from artifacts induced by thresholded evaluation?
- What internal observables best track capability transitions?

11.4 Thermodynamic-budget conjecture

The upper bound on a learning system's adaptive performance depends jointly on data, model capacity, time budget, and physical dissipation budget.

Research questions

- Can generalization, memory formation, and energy dissipation be brought into a common formalism?
- Are there experimentally meaningful lower bounds on the cost of adapting to classes of environments?
- Which architectural motifs improve adaptation per unit of physical cost?

A fifth, more speculative question cuts across all four conjectures: can a theory of intelligence be built around effective coarse-grainings of dynamics rather than around functions in isolation? If so, the right mathematical language may involve not only probability and control, but also symmetries, renormalization-like ideas, and structure-preserving maps between internal and external dynamics.

12. Conclusion

The central intuition behind this essay can now be stated more cleanly than at the start. 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, these systems become regulators. Under sufficient complexity, memory, and multiscale feedback, they begin to exhibit behaviors we naturally describe as intelligent.

This remains a working agenda, not a finished doctrine. Several claims in this paper are offered as conjectures, not settled facts. But that is precisely the point. The goal is to put forward a picture that is strong enough to organize research, weak enough to remain falsifiable, and concrete enough to generate formal questions. If even part of this picture is right, then a "physics of AI" is not just a slogan. It is a legitimate scientific program linking machine learning practice, general intelligence, and safety.

Acknowledgments

I am especially grateful to Ziming Liu for publicly articulating and motivating the possibility of a "physics of AI." His ongoing reflections have been a major source of intellectual encouragement for pursuing this line of thought. More broadly, this essay draws inspiration from the work of W. Ross Ashby, Roger C. Conant, David Marr, Naftali Tishby, William Bialek, Rolf Landauer, Sergey Levine, Karl Friston, Patrick McMillen, Michael Levin, Jonathan Richens, Tom Everitt, David Abel, and many others whose theoretical contributions make this discussion possible.

Author contributions

The initial motivation, core intuitions, and problem framing were proposed by Macheng Shen, including the emphasis on dynamical systems, the idea that training may transfer effective environmental dynamics onto a new substrate, the role of multiscale feedback, and the connection to AI safety. Conceptual restructuring, theoretical cross-linking, argument tightening, literature integration, and drafting of the present English working-paper version were developed

collaboratively with GPT-5.2 Pro. This document is intended as a public perspective essay designed to invite discussion, criticism, and further development.

References

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.

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.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. 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., Everitt, T., & Abel, D. (2025). General agents need world models. In *Proceedings of the 42nd International Conference on Machine Learning* (PMLR 267, pp. 51659-51687).

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems 36*.

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

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

Zhang, Z., & Xu, F. (2024). An overview of the free energy principle and related research. *Neural Computation, 36*(5), 963-1021.