Superintelligence from Complexity

Learning from the gap between human prediction and reality

Mikhail Gorelkin
Independent Researcher
Principal AI Scientist & Principal AI Systems Architect
magorelkin@gmail.com

This is a concept paper. It presents an argument and a proposed direction, not a completed research contribution. It is posted here to establish a dated record of the framing and to invite engagement; a more developed treatment, with full engagement of the relevant literature, is planned as a follow-up.

Abstract

Modern AI systems are trained primarily on the accumulated symbolic output of human civilization — a map of what humans have thought, discovered, and argued. This paper proposes that a complementary training signal, currently unexploited at scale, lies in the systematic divergence between human predictions and actual outcomes. Public discourse continuously generates forecasts about politics, economics, conflict, technology, institutions, and social change, yet these forecasts are almost never extracted, formalized, scored, and analyzed as a learning resource. Treating this planetary archive of prediction-reality gaps as a developmental environment may provide a distinct source of pressure on model capability — one that scales, that exposes systems to the structural features of real complexity (non-stationarity, adversarial adaptation, delayed feedback, reflexivity), and that teaches not only what humans believe but how human models of complex reality fail. The paper sketches the proposal, identifies the principal technical and methodological difficulties, and situates the argument against existing work on scaling, agentic training, and forecasting.

Keywords: superintelligence, developmental AI, forecasting, prediction-reality gap, agentic systems, complexity, training regimes

---

In 2021, the Federal Reserve, the Treasury, and a broad consensus of mainstream economists predicted that the inflation surge would be transitory — a short-lived byproduct of pandemic disruption that would fade as supply chains normalized. It didn't. The models behind those forecasts were not stupid. They weighted historical inflation dynamics, output gaps, and the recent decades of stable prices. They were wrong in an instructive way — underweighting supply-chain persistence, labor-market shifts, and the feedback between expectations and wage-price behavior. The prediction failed, discourse moved on, and almost nothing was done to systematically extract what the failure revealed about the models that produced it.

This happens thousands of times a year. Public prediction — about elections, wars, markets, technologies, institutions, social movements — is produced at enormous volume and audited almost never. Humanity generates a vast stream of forecasts without converting that stream into a disciplined learning system. Predictions shape decisions, circulate through institutions, influence markets and public opinion, and then disappear into the archive. Rarely are they extracted,

formalized, tracked, scored, and analyzed across time and domains.

This may be one of the richest untapped training environments for advanced AI.

The overlooked training signal

Modern AI is trained primarily on the accumulated symbolic output of human civilization: books, papers, code, documentation, forums, articles, arguments, explanations. The corpus contains an enormous map of what humans have thought, discovered, debated, misunderstood, and built.

But the corpus does not contain reality in the same way that lived engagement with reality does. It contains human descriptions, interpretations, narratives, and predictions about reality. Some are accurate. Many are partial. Many are wrong. And some are wrong in structurally revealing ways.

The important signal is not only what humans said. It is the difference between what humans expected and what actually happened.

Predictions as compressed causal models

Every prediction is a compressed model. When a political analyst says a regime is stable, when an economist says inflation will fall, when a military expert says a conflict will resolve quickly, or when a technology commentator says a product will dominate, they are revealing an implicit model of causality — assumptions about incentives, institutions, resources, social mood, adversarial behavior, second-order effects, timing, feedback loops, hidden constraints. Usually these assumptions are not fully explicit. They are folded into narrative.

When the prediction fails, the failure is not just an error. It is evidence about the limits of the underlying model. What was missed? Which variable was overestimated? Which feedback loop was ignored? Which actor adapted? Which causal frame was wrong from the beginning?

A sufficiently capable AI system could turn this into a planetary-scale learning process.

The proposal

Imagine an agentic system that continuously reads newspapers, television transcripts, expert reports, geopolitical analyses, financial commentary, public speeches, think-tank publications, and institutional statements from across the world. Its task is not to summarize. Its task is to extract predictions, formalize them into structured claims, attach time horizons and confidence where possible, track outcomes, compare forecasted trajectories with actual ones, and analyze the structure of error.

Over time, it would learn not only what people believe about politics, economics, conflict, technology, institutions, and social change. It would learn how human models of complex reality fail — the recurring structural reasons why confident expert models break against contact with events.

This is training not on text alone, but on the gap between text and reality.

Why this is harder than it sounds

The proposal has several hard problems, and any serious version of the argument has to acknowledge them.

Extraction is non-trivial. Many of the most consequential forecasts are not cleanly stated. They are hedged, implicit, embedded in strategic narrative, or phrased to be defensible against any outcome. Identifying what was actually claimed — and what would count as confirmation or failure — is itself a significant inference problem. Forecasting tournaments exist partly because operationalizing prediction in the wild is so hard.

The corpus is not a random sample. Media-archive predictions skew toward events that got discussed, toward confident claims (hedged ones do not make headlines), toward domains the media covers, and toward the institutional and linguistic contexts that produced the archive. A system trained on this learns the failure modes of that distribution, not of human cognition in general. This has to be corrected for, not wished away.

The system is grading its own homework. If extraction, formalization, and outcome assessment all run through the system's own judgment, the feedback loop is reflexive. This is not fatal — humans face the same problem and still learn — but it means the system's blind spots will be systematically invisible to it in exactly the places where they matter most. External anchoring, adversarial review, and held-out human scoring have to be structural features of the design, not afterthoughts.

Post-hoc is not prospective. Learning why documented forecasts failed is not automatically the same as navigating novel situations where no record yet exists. The bet is that enough exposure to varieties of model failure produces better priors about where current models are likely to break — the same mechanism by which human superforecasters improve by studying past errors. The bet is defensible, but it is a bet.

None of these objections defeats the proposal. But a serious version of it is an architecture, not just a dataset.

Why this is different from prediction markets

Prediction markets and forecasting tournaments already try to improve human prediction. They are valuable but limited. They depend on explicit questions, bounded outcomes, and participants who know they are forecasting.

The global media and institutional archive is different. It contains billions of implicit and explicit forecasts that were not designed for scoring. Most consequential predictions are embedded in ordinary discourse: policy statements, strategic narratives, investor memos, expert interviews, military assessments, ideological claims, public explanations of unfolding events.

This messiness is exactly why the environment is valuable. It is non-stationary, adversarial, reflexive, and full of delayed feedback. In other words, it has the structure of real complexity.

What such a system would need to do

A serious version of this idea requires several layers of capability.

Prediction extraction. Identifying explicit and implicit forecasts across languages, media formats, institutions, and rhetorical styles.

Formalization. Translating vague claims into structured propositions: what is expected to happen, under what conditions, by what time, with what implied probability, according to which causal model.

Outcome tracking. Connecting forecasts to later events while handling ambiguity, partial fulfillment, moving goalposts, and contested interpretations.

Error analysis. Classifying why predictions failed: missing variables, incorrect causal assumptions, ignored feedback loops, adversarial adaptation, ideological distortion, overfitting to recent events, institutional incentives, regime change.

Model improvement. Using these errors not merely to score forecasters, but to build better representations of complex systems.

Humans cannot do this at planetary scale. The volume is too large, the horizons too long, the domains too heterogeneous, the feedback too entangled. Institutions also have weak incentives to preserve and score their own failed predictions; public discourse moves forward by narrative replacement rather than systematic correction.

This is why the opportunity is significant. The world already produces the data. What is missing is the machinery that turns the data into developmental feedback.

The deeper hypothesis

Superintelligence may not emerge only from scaling the imitation of human outputs. It may require scaling confrontation with the conditions under which human models fail. Human civilization continuously generates hypotheses about its own future. Most are never audited. From the perspective of AI, this unaudited stream is not noise. It is an enormous experimental record of minds attempting to model complex reality and being corrected by events.

If artificial intelligence can learn from that correction more systematically than humans can, it may develop a form of intelligence that is not just broader than human knowledge but structurally better at recognizing when human models are about to break.

Conclusion

The next leap in AI may not come only from larger models or more static data. It may come from systems trained against reality itself — systems that extract human predictions, compare them with actual outcomes, and learn from the recurring failure of our models.

Humanity already produces the raw material for this process every day. We call it news, analysis, commentary, strategy, policy, public debate. But underneath those forms lies something deeper: a planetary archive of predictions colliding with reality.

Superintelligence from complexity means learning from that collision.

This paper follows from an earlier essay by the author on developmental pressure as a training regime for advanced AI systems. The present work proposes a concrete source of such pressure; a more developed technical treatment is planned as a follow-up.