A MYTH FOR AGI

Lian Sidorov

From a number of perspectives, 2025 was a watershed moment in artificial intelligence. For the first time, after years of galloping successes in large language model (LLM) scaling strategies, there was now widespread acknowledgment that LLM approaches alone would probably not be able to take us all the way to artificial general intelligence (AGI), and that world model building was an essential step toward that goal [1, 2]. More importantly yet – Geoffrey Hinton, freshly anointed with the Nobel Prize in physics for his contributions to neural networks and deep learning, came out to forcefully argue that the only way to remain in control of AGI and avoid human extinction at the hands of this new apex species was to build maternal instincts into any future superintelligence [3] – in other words, that the only type of control we could hope for was motivational, not technical [4].

As we navigate the present moment, with its convergence of multiple global crises and unstoppable AI acceleration, it is often difficult to discern our true movement direction through the confusion of corporate hype, starry-eyed talking points about a post-scarcity utopia and universal basic income (in ironic defiance of all present policies and tax structures), countless doomsday scenarios involving malevolent or inept AI, cyberwarfare, the steady labor market erosion as a result of AI deployment, or the spectacular scientific advances achieved by narrow AI like AlphaFold. But without trying to step back from this constant drumbeat and understand our current evolutionary vector, we risk losing even more control over a process that is rapidly propelling us into an almost unrecognizable future.

To take such a step back, we propose a series of discussions which, in our opinion, address some of the fundamental issues underlying this transitional moment for our species:

1. The fact that with the imminent arrival of AGI and brain computer interfaces there are a number of key questions we no longer have the luxury to sweep under the carpet – questions about the boundaries of consciousness and self, about individual versus societal and civilizational goals, about control over the narrative of science and our own place on the spectrum of intelligent agents in the universe, soon to include a superintelligence of our own creation.

2. The fundamental question of intelligence evolutionary dynamics – understanding the drives that shape the emergence and evolution of different cognitive agents, from the first molecular

networks learning to navigate their environment to human societies and increasingly more autonomous artificial intelligences, including the possibility of a future AGI civilization

3. The distinction between first and second-order intelligence – where first order intelligence refers to an individual's priorities in order to survive and thrive, while second order intelligence is the ability to place one's goals in the context of the next level of organization – be it cell-to-tissue morphogenetic development, or human-to-society relative priorities.

4. Addressing the question of AI alignment and AI safety from a somewhat unorthodox perspective – not from the tight grip of a master-slave relationship, which seems to be the way we are currently constructing most scenarios, but working backwards from the perspective of ASI (artificial superintelligence) and in particular the concept of *ikigai* (meaning, reason for living) introduced by Yampolskiy in the context of AI alignment [5] . However, while Yampolskiy focuses on human *ikigai* in a post-AGI, post-work, post-scarcity future, it could be argued that the issue of ASI-level *ikigai* is far more critical to explore if we wish to engineer a safe, symbiotic relationship with this new apex intelligence, to find our own value proposition in that partnership and start designing the genetic building blocks of AI world models in which such existential questions can be successfully, constructively answered for both us and our synthetic successors.

These are not obvious questions, yet our hypothesis is that they will become increasingly more relevant as we move into a wildly turbulent transition period for humanity: the AGI alignment problem cannot be separated from our own goals as a civilization, and how well we have managed to align with those stated goals remains a point of contention. Is AGI a threat to human civilization, or the only way to save it? While no one denies that there are thousands of scenarios in which AGI could precipitate human extinction, an equally important question is whether there is a single plausible scenario in which humanity survives and thrives, from its present condition, in the absence of some radical self-transformation demanded by the emergence of AGI. This is not rhetorical exasperation with over 10,000 years of failed political experiments, but a simple question in the face of this incontrovertible evidence: does the human species, in its present form, have what it takes to successfully execute the civilizational program it has been trying to codify for millennia through its system of laws and social charters, its art and science – or is there an intrinsic hardware limitation in our ability to balance first- and second-order intelligence goals?

Fortunately, there are researchers who have already cut long tracks into this abstract space, and our hope is to engage with their findings as we tackle these questions over the coming years: AGI pioneers like Ben Goertzel, computational biologists like Michael Levin, topological geometrodynamics theoretical physicist Matti Pitkanen, and Roger Nelson, who developed the Global Consciousness Project at the Princeton Engineering Anomalies Research Lab and has

been running it since 1998, have for decades explored the far reaches of our common intellectual space and patiently mapped out some of the processes and laws they have observed at the boundary between matter and intelligence in various scales and substrates, starting to model such dynamics and speculate about the path that lies ahead of us, once we dare to let go of our anthropocentric lens and view reality from a more balanced perspective.

And from that perspective, the question perhaps is no longer whether man – man in his present biological form, with all his known moral capacity for good and evil – will remain the species in charge of our cultural legacy, our scientific aspirations and our envisioned ethical architecture; the question is whether the AI successor we manage to build over the next few decades can embody, protect and prioritize these great civilizational achievements of humanity to a better degree than we did. If we can engineer such a model, if we can ensure that it wins what will undoubtedly be an incomprehensibly fast evolutionary race against other AI species, then the question of mankind's survival may be resolved as part of that equilibrium: because if these are the cultural artifacts valued by AI, their creative source may remain forever partially shrouded in something that cannot be fully coded, some Godel-type aperture on universal truths, and one does not destroy what one may still need at some future time – just like we try to protect yet undiscovered species in the Amazon for their potential pharmacological benefit.

The recent realization that LLMs alone will not take us to AGI and that building formal world models (i.e. a neurosymbolic approach) is critical, both buys us additional time and orients us toward the right strategy when it comes to AI alignment. But static world models will probably be insufficient as a general framework, because of AGI's extraordinary rate of self-modification once allowed to recode itself. As Ben Goertzel observes in [6], one needs dynamic , evolutionary models as part of our AGI strategy. It is obvious from our own history that as capabilities grow, our goals evolve: with future superintelligence expected to evolve at a staggering rate, how might the arc of that evolution be bent in a positive direction – is there anything we can engineer into the embryo of a future ASI that would maintain that trajectory despite countless unimaginable modifications over the course of its history?

It is here that the work of researchers like Michael Levin (who studies emergent agentic behavior across biological and synthetic substrates, from anthrobots/ xenobots and simple computer programs like Bubble Sort to morphogenetic behavior in regenerating tissues) can give us a hint about the patterns to follow in the evolution of intelligence goals across different scales. Beyond mere survival and multiplication, what other universal drives and goals are at play? Do they recur in different systems and scales? Are they modifiable by external influences? Do they remain consistent over generations?

The fact that Levin has recently resurrected the strangely anachronistic notion of Platonic Forms [7] as "attractors" in the space of embodied intelligence manifestations, whether this be morphogenetic goals like a particular limb shape in a regenerating organism, or a mathematical principle we discover, is very useful in this context. Because the evolution of an agentic system

is not just a matter of starting conditions and environmental pressures, not just a matter of survival, but one of goals. This applies to humanity, it applies to the AI systems we build today and task with various functions, and it will apply to the ASI that AGI builds, far beyond our understanding and control.

And so the fundamental question we may need to confront is this: can we design a goal so compelling, so difficult to reach, so binding to an entire AI civilization, that it prevents ASI evolutionary drift while keeping it aligned with human values, and keeps humanity an essential part of that endeavor?

Can we create a myth for AGI?

References

1.Marcus, Gary Three years on, ChatGPT still isn't what it was cracked up to be – and it probably never will be. A skeptic's pre-mortem Nov 29, 2025

2.Marcus Gary Neurosymbolic AI is quietly winning. Here's what that means – and why it took so long July 13, 2025

3.Geoffrey Hinton vs. The End of the World. Globe and Mail Oct 7, 2025

4.Yampolskiy, R. V. (2024). *AI: Unexplainable, unpredictable, uncontrollable*. Chapman and Hall/CRC.

5. Ziesche, S., & Yampolskiy, R. V. (2025). *Considerations on the AI Endgame: Ethics, Risks, and Computational Frameworks*. Chapman and Hall/CRC.

6. The Truth About Our AI Future w/ Ben Goertzel

7. Levin Michael 2025 Symposium on the Platonic Space