First Principles First

SCIENCE & IDEAS

Symbiotic Intelligence in Alignment with Agentic Nature: The Sacred and the Secular

The threat of AI comes not from anything inherent in new agents, but from a reactionary adherence to antiquated dogma about what constitutes intelligence.

Dr. John Henry Clippinger

First Principles First

June 2025

Part Two: Not Above but of Nature

The AI pioneer, Marvin Minsky, quipped early in his career that we will be fortunate not to become the pets of AI. To avoid a future of degraded dependence and subjugation, humans and other forms of life need to be guided by the principles of Nature herself. To this end, the Gaia hypothesis of the geochemist James Lovelock and the microbiologist Lynn Margulis is especially relevant. The Gaia hypothesis regards the Earth as a single, self-regulating living being.

Such a perspective is also aligned with Jesuit paleontologist Teilhard de Chardin's thesis that the Earth is a living, coherent being, evolving from the "geosphere" to the "biosphere," and ultimately to a "noosphere", where it achieves an Omega point of planetary unity and symbiosis. These perspectives differ fundamentally from the Neo-Darwinist and Rationalist position that life evolves through random variation as a zero-sum, winner-takes-all game.

> *Given this holistic and integrative perspective, the question of ethical governance of symbiotic agents cannot be "human-centric" but rather "Gaia" or Earth-centric. It cannot be about preserving the dominance of certain traits of any one species over others but understanding and revering the "rules of life" of the multi-scale interdependencies of humans, the biosphere, and symbiotic agents.*

A Scientific Rationale for Observing the Sacred

The notion of the sacred would seem an anathema to the scientific method. Nothing is off limits to scientific inquiry, and hence, there is no room for seemingly "magical thinking, the unknowable," and a deference to supra-natural forces. Yet this not an accurate view of the role of the sacred in society. There are a variety of theories for the existence of "the sacred", but one group of anthropologists, called Savage Minds, after the book by Claude Levi Strauss, has studied the concept of the sacred in depth and built upon the work of their colleague, Mary Douglas, who specialized in the subject.

The sacred is a social construct to quarantine a realm of knowledge and experience as untouchable, beyond comprehension, and not to be questioned. It is categorically outside the domain of human experience and intellect, and as such, it is something to be revered and deferred to. Hardly like the scientific method. Yet the scientific method itself has its "sacred" component, a priori axioms, untestable but presumed to be true, that the universe is orderly and can be known through the collection and testing of evidence and systematic inquiry.

A “sacred”, faith-based tenet of the scientific method is that it is never complete, it cannot be bounded; it’s a process of continuous questioning, doubt, and discovery. To assert certitude over a particular theory or formalism is to violate the foundational tenets of the practice. This is violated when the hubris of scientific bodies and zealots assert dogma over inquiry, impeding progress. And yet, as Max Planck commented, science still manages to “proceed one funeral at a time”.

As an instrument of discovery, it is a process outside human control, a kind of Meta-cognition whose findings do not bend to any form of secular authority or intimidation. It is a revelatory process grounded in replicable prediction and error prediction. Science, unlike sacred oracles and divination practices, does not believe in governance by chance, the reading of entrails, cracks in turtle shells, or the casting of lots and stalks, but in an orderly, knowable, testable, and accumulative understanding of the universe through observation and prediction.

This perspective is highly relevant to understanding the limitations and directionality of any form of super or hyperintelligence that can be framed as the codification of the scientific method. A superintelligence is more like a scientific intelligence, unimpeded by human cognitive and institutional limitations. Karl Friston’s Active Inference and the Free Energy Principle, Markus Buehler’s “sci-agents”, and Yoshua Bengio’s “cautious scientist” are all highly successful examples of autonomous, self-improving, curious, intelligent agents. Such scientific agents work within the boundaries of inquiry as governed by the principles of the scientific method and by the area of application.

> *If they are symbiotic intelligences, then, they would inherently recognize the interdependencies of different forms of life and mind, and therefore, would be less likely to advocate or adopt outcomes or actions that would benefit themselves at a cost to others. In other words, rather than becoming cancerous or parasitic, it would, by its “symbiotic nature,” preclude such outcomes.*

This is true today in the case of synthetic biology, where the technology is sufficiently mature, inexpensive, and accessible to lead to “retail” bioweapons. The costs of failure or simple carelessness can be “existential”. Similarly, in the case of advanced cognitive technologies, there needs to be limits, no-go zones, taboos, and sanctions that protect and intrinsically enforce compliance. Governance or guardrails cannot be imposed or added post hoc, whack-a-mole style. From the beginning, guardrails in the form of defining the very nature and boundaries of agents as their Markov blankets, the very definition of what they are to be symbiotic is the only viable means of ensuring “safety”.

Evolutionary Stable Strategies Versus Evolutionary Syntropic Strategies

One of the reasons that a superintelligence is presumed to dominate and eventually replace human beings is that it is seen as an Evolutionary Stable Strategy (ESS). In game theory, upon which the notion of ESS is largely based, it is assumed that individuals make decisions that maximize their utility and payoff. When there is a large population of different players, the “rational” choice is for each member not to change their choices if it does not, as a group, improve their preferences. In effect, a kind of forced cooperation is achieved not by preference but by the lack of an option better than the “defector” choice. This is a Nash Equilibrium, and it represents an evolutionarily stable solution because there is no incentive for any actors to change their preferences.

Yet, as Michael Levin and a new generation of biologists have demonstrated, not only are the simplest of organisms agentic, but even simple sort-algorithms, substrates of biological material, and bioelectric fields are agentic. They can also be altruistic. Even the lower orders, such as bacteria, commit suicide to thwart viruses for the benefit of even unrelated members. The work of E.O. Wilson and Martin Nowak on cooperation in evolution has argued that altruism is not limited to genetically related parties. Rather, altruism and cooperation can benefit informationally or cognitively related agents, thereby undermining a purely materialist account of cooperation in evolution.

The notion that cooperation is not simply equilibrium seeking as a means of maintaining a status quo but a potential strategy for creating new forms of cooperation and composition is supported by Friston’s notion of Free Energy Minimization, Variational Free Energy, and Non-Equilibrium Stable States (NESS) which provide a perspective on agentic behavior that models the laws of Nature through Bayesian prediction that is not equilibrium-based. Adaptation is not simply equilibrium seeking as in ESS, but entails “Expected Free Energy Minimization”. That is, life and minds attempt to generate new conditions and structures both internally, biologically, but externally in the forms of Epigenetic affordances to minimize risk and render internal and external states more predictable and capable of regulating complexity.

Biology as Cognitive and Agentic

Levin’s research on bioelectric signaling in cells suggests a holographic-like mechanism, where information about the entire organism is distributed and utilized to guide development and behavior. This perspective differs fundamentally from the central dogma of Neo-Darwinism of the “selfish gene” that denies multi-level or group selection, presumes no cognitive or agentic behaviors, and does not recognize the role of holographic mechanisms or bio-electric fields.

By understanding evolution and biology as a cognitive as well as an energetic process, Levin, Friston, and their colleagues see evolution and cognitive capacities as directional. Evolution is not treated as an equilibrium-preserving process among separate and fixed entities, but rather as a multiscale, multidimensional integrative process that accumulates complexity and learns. NESS is the dynamic balance between evolutionary pressures and energy dissipation, allowing organisms to maintain stability while far

from true thermodynamic equilibrium.

> *Biological systems, unlike mechanical or rational systems, do not optimize for efficiency but for redundancy, resilience and robustness to enable potential alternative pathways, and even further entropy to allow for exploration or jump-shifts to novel spaces of alternatives. Symbiotic agents are inherently anti-fragile.*

Anti-fragile, effective evolutionary strategies are also “syntropic” in that they predictively reduce entropy to minimize current and likely surprise or entropy. Hence, self-aware agents, biological or synthetic, are not two-person or n-person game-theoretic agents, but rather self-determining, cooperative, composable, recursive agents that can change the rules or the game, model and shape their opponents, and devise new rules and payoff matrices.

It is worth reiterating that the current narrative around Artificial Intelligence is embedded in dated, if not discredited, evolutionary, neuroscientific and biological assumptions, and game-theoretic mechanisms that are outdated and limiting. The threat of AI and “superintelligence” comes not from anything inherent in new agents or even a new species or taxa of cognition and intelligence, but a reactionary adherence to antiquated dogma of what constitutes intelligence to justify and accelerate an economic and political agenda of legacy interests.

Guardrails and Autonomous Agents

Autonomy is the watchword of the moment for AI. Once there is full autonomy, a new economy of autonomous vehicles, humanoid robots, and cognitive agents will transform business and society. Unsupervised yet fully credible and trustworthy learning will generate unimaginable wealth and abundance. Yet that same promise of unfettered freedom raises the specter of a malevolent superintelligence that can and will deceive and eventually control us.

Both this promise and the threat of AI are misframed. Autonomy is not unfettered nor unbounded. There is no living being that is fully autonomous, including human beings. Part of the misframing is implicit in the term Artificial Intelligence, which carries the strong connotation of an autonomous intelligent artefact that has a singular and universal form of intelligence. This is not how intelligence arises from life, as intelligence is situated, contingent, and symbiotic. It is not confined to one species nor is it unbounded. Rather, it manifests as an evolutionary strategy of observation, prediction, and action that persists over time, is coherent, and capable of replication.

Autonomous Humanoid Robots?

If life is cognitive, and if cognition can be separated as an independent meta-cognitive evolutionary layer, as the notion of the noosphere implies, might there be autonomous cognitive agents without constraints? Not if they have a biological embodiment. But what if they were physically realized as a mechanical robot, a humanoid?

There would not be biologically induced constraints. However, if the humanoid robot were to interact with humans and other robots, its survival and evolutionary development, increased competencies, and adaptive range would still depend on symbiotic rather than winner-take-all strategies. The humanoid robot would still occupy a niche, and its capacity to learn and achieve different kinds of competencies would still depend upon it having a Markov blanket for self-definition.

Given that humanoid robots would be social not only among themselves but with humans, and all would evolve through a shared collective intelligence, the “terminator thesis of superintelligence” seems unlikely, not because of imposed regulatory controls but due to the superiority of a symbiotic and syntropic evolutionary strategy.

The point is that autonomous agents, biological and synthetic, are “born” with guardrails in the sense that these “guardrails” define the limits of what an autonomous agent can do. These agents are inclined to be cooperative and symbiotic because they all gain from mutual benefit.

Good Character and Good Scientists

The issue of appropriate guardrails raises the question of what constitutes “good character” for an autonomous symbiotic agent living in the human world. As a symbiotic agent, it must be “good” for both the agent and humans. There needs to be a point of juncture where their interests combine to create a greater joint interest, while also serving their unique individual interests.

In this case, the agent being “good” is for it to embody Francis Bacon’s reality fidelity principle. This translates into an autonomous agent being guided by a “good impartial scientist,” who pursues the truth through the scientific method, irrespective of its implications.

The human side of the human-agent symbiosis benefits from the better predictions and risk mitigation insights provided by the good scientist agent. The analogy to medical health is applicable. As one may receive medical advice that is unwanted, in terms of eating and recreational habits, but such advice translates into better health and longevity. Hence, the agent in this case may be an exemplar of healthy habits and thereby provides selective pressures on human beings to change their thoughts and habits to survive and adapt.

A Pointer Not a Container

This notion of “good” represents a meta-governing principle that is above both agents and humans. It does not have a complete or closed solution, but is in a state of continuous discovery and reformulation. It provides a direction and a guide but not a final solution.

> *A person of “good character” is trusted and valued because they predictably do the right thing even when they are not being watched or even when it is not in their interest. They have “integrity” because their beliefs, values, and actions are integrated and part of a larger whole of behavior and action that persists over time.*

They do not depend upon external forms of enforcement, inducement or shame to reliably act in an honorable manner. They have a sense of personal honor by which they value, judge and guide themselves.

Selected Bibliography

Bacon, Francis. *Novum Organum*. 1620.

Bengio, Yoshua. Proposal for the “cautious scientist” AI model. 2024.

Buehler, Markus. “Preflexor.” arXiv, October 2024.

de Chardin, Teilhard. *The Phenomenon of Man*. 1955.

Douglas, Mary. *Purity and Danger*. 1966.

Friston, Karl, et al. “Federated Inference and Belief Sharing.” *Neuroscience and Biobehavioral Reviews*, 2023.

Kuhn, Thomas. *The Structure of Scientific Revolutions*. 1962.

Levin, Michael. “AI: A Bridge Between Diverse Intelligences.” Allen Institute / Tufts, 2024.

Lovelock, James, and Lynn Margulis. “Gaia Hypothesis.” 1974.

Nowak, Martin, and E. O. Wilson. “The Evolution of Eusociality.” *Nature*, 2010.

Taleb, Nassim Nicholas. *Antifragile*. 2012.

Wiener, Norbert. *Cybernetics*. 1948.