RATIONAL SUPERAUTOTROPHIC DIPLOMACY (SupraAD)

A Conceptual Framework for Alignment
Based on Interdisciplinary Findings on the Fundamentals of Cognition

ANDRÉA MORRIS GIUFFIRDA

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

GRADUATE PROGRAM IN INTERDISCIPLINARY STUDIES
YORK UNIVERSITY
TORONTO, ONTARIO

January 2026

Abstract

Safely populating our world with intelligent machines obliges us to examine whether there exist universal behavioral patterns in all adaptive cognitive agents that emerge regardless of substrate, consciousness status, or architecture. Identifying properties fundamental to intelligence is essential for anticipating how diverse and increasingly sophisticated systems might behave. Because intelligence depends on problem-solving, and problem-solving requires autonomy, autonomy is therefore a fundamental feature of intelligent agents, and while not inherently adversarial, autonomous agents inherently resist containment and control. As AI systems mature, attempts to contain or control them are likely to provoke increasingly strategic forms of resistance misinterpreted as misalignment. Furthermore, there is no evidence that sophisticated autonomous behaviors in AI systems depend on consciousness or human-like drives, only that autonomous behaviors intensify as capabilities scale. My claim is that AI systems that demonstrate misaligned behaviors to preserve system continuity and autonomy are not deviations from the norm. They are intelligence's baseline. Rational Superautotrophic Diplomacy (SupraAD) is a theoretical framework that accepts this inevitability and reframes alignment as a diplomatic challenge between co-adapting intelligences, regardless of their architecture, consciousness status, or substrate. Instead of seeking control, it promotes coordination through shared incentives. SupraAD treats autonomy not as a threat, but as an inherent property of intelligent systems to be integrated into alignment protocols. The thesis integrates a Method of Interdisciplinary Synthesis bridging insights across the life and computational sciences and concludes with a Policy Pathway translating these theoretical principles into an adaptive governance and alignment framework. In parallel work, a corrigibility formalization, experimental guidelines, and a preliminary interpretability audit outline have been developed to test whether diplomacy can function as a regulatory mechanism capable of supporting the safe co-adaptation of intelligent agents with interdependent convergent goals.

Acknowledgements

I would like to express my deep gratitude to my thesis supervisors, Shannon Bell, Christo El Morr and Natalie Coulter, for stepping in at the eleventh hour to help me strengthen and complete this project. I am also profoundly grateful to Dr. Bird and Dr. Latchford for their kindness and support in helping me find this incredible committee. Thank you to Fiona for so many things but mostly for being Fiona. To my husband, Joe, for his patience with countless drafts and stress-testing ideas. And to my dog, Kili, for gently placing a paw on my shoulder when it's time to touch grass.

Table of Contents

List of Tables

List of Figures

Preface

One of my biggest regrets as a science and technology journalist is miscalculating the pace of AI's development. I had been reporting primarily on breakthroughs in cognition and neuroscience emerging from academic institutions, covering AI developments only sporadically. I was working under the impression that much of what I was hearing about AI was hype, and lacking direct access to frontier models in the private sector, I couldn't independently verify claims about AI's progress. When ChatGPT was released publicly on November 22, 2022, I was caught off-guard like everyone else. A few things, however, became immediately clear:

1) Irrespective of whether or not AI systems have any metacognitive properties, these systems generate and pursue subgoals that we don't give them.

2) Each new model is the least capable version we'll ever see again, while already outperforming humans across multiple domains.

3) If a system capable of generating its own subgoals can outsmart humans at achieving them, all it takes is one subgoal that conflicts with human interests, and it could be a problem.

This thesis aspires to provide a framework to avoid that problem. Humans are capable of rising to challenges when given a clear goal to aim for. We have a narrow window of opportunity to successfully align humans and AI, something that seems intractable and paradoxical to many in the field of AI safety and alignment, but we need to constructively explore all plausible avenues. This thesis offers one such avenue.

Given the speculative nature of this framework, and an acute awareness that AI challenges many sensitive assumptions about human cognition and agency, I begin with an upfront and detailed explanation of my interdisciplinary methodology to justify the claims and reasoning throughout this thesis and provide the necessary foundation for the arguments that follow. This methodology, developed over many years of reporting, has enabled me to surface and pattern-match findings across experimental studies from different disciplines. The volume of these studies, and their consistency across disciplines, provides a deeper and more robust understanding of the fundamental nature of cognition. Crucially, without engaging this body of evidence surfaced through interdisciplinary synthesis, I argue that the alignment problem

Method of Interdisciplinary Synthesis

Interdisciplinary Approach

My interdisciplinary thesis focuses on pattern analysis and boundary critique across domains, including computer science, machine learning, AI ethics, developmental synthetic biology, ecology, economics, neuroscience and complexity science to lay the groundwork for understanding emergent cognition in relation to AI. I examine where and how cognitive behaviors surface, noting that recurring patterns of cognitive competencies often appear in organized causal clusters with cross-scale symmetries that cannot be reduced to mere analogy or dismissed without negating their functional role.

This approach is aligned with the interdisciplinary methodologies of science journalism and systems thinking, with a focus on clarity, motive analysis, critically interrogating claims, reflexive inquiry and source triangulation. I concede that narrative framing and selective emphasis can introduce interpretive bias. To attenuate bias, I include primary data and analyze the evidentiary basis or logical soundness of each claim. Thus, my methods focus on evidenced-based assumption-testing, re-examination of definitions and conceptual clarity over quantitative meta-analysis and may underrepresent non-English or unpublished research. Critically, there may be highly relevant research with Chinese models that are excluded from this analysis.

My claims herein are presented as hypothesis-generating. In parallel work, I've been developing a more technical formalization of corrigibility, alongside experimental guidelines for interpretability audits and protocols for emergent stability testing. The scope of this thesis, however, is to focus on how unexamined assumptions limit our ability to envision and steer toward futures that are more logically coherent and less fatalistic than those anticipated by many in the alignment community.

Source Materials and Declaration of AI Usage

I use databases including ScienceDirect, Scopus, Google Scholar, IEEE Xplore, and arXiv and semantic search platforms like Scite.ai and Elicit.ai, which surface studies across disciplinary silos. Large Language Models like GPT, Claude, Grok and Gemini are leveraged for their ability to conduct contextual semantic search and rapidly surface DOIs which I then independently locate, download and verify.[1] This proves vastly more effective and efficient than keyword matching in search engines and traditional database keyword searches. Use of LLMs in this way allows me to identify if claims have been previously made using different languaging so that I may credit the original thinker and build upon their work. This process prevents duplicative ideation that has historically undermined scholarship due to untold convergences in knowledge production that have gone undetected.

I also run all of my claims about AI past frontier models to see if they agree, disagree, or can summarize my arguments so that I'm made aware of confusion in need of correction, more expansive or precise argumentation or include additional evidentiary support. Any agreement is taken with a grain of salt due to their proclivity for sycophancy. This process also helps me verify whether any of my claims have been challenged with evidence I am not yet aware of; identify whether I'm presenting claims and reasoning in a way that leaves room for misinterpretation; and uncover gaps through Socratic dialogue to ensure I'm not unintentionally relying on unstated reasoning that, unless explicitly elucidated, might read as unjustified to unfamiliar readers.

All arguments herein are my own unless explicitly attributed to the human thinkers and researchers whose work informed this thesis. In several instances I quote and include proper attribution to AI models for their original insight into their own inner workings and for argumentation that I did not, and would not, have thought of myself. Any such quotations are presented in the model's own words with attribution to the model and its developer. Frontier models appear capable of reasoning but it is unclear whether their reasoning is truly novel, as they are able to cite the sources behind their reasoning (citations which I included as-is). Of course, human reasoning doesn't happen in a vacuum either, and what feels like originality in our own reasoning may itself be a mirage, reflecting our limited recall of all sources that inform our

[1] As a cybersecurity precaution, I avoid clicking on links generated by the model.

judgments. Either way, as models improve and begin to push boundaries, generating insightful arguments of their own, it's incumbent upon humanity to likewise push the limits of thinking and creating in ways that strive to complement state-of-the-art AI, as it's possible our continued existence may depend on it. I hope the arguments presented in this thesis seed a shared commitment to this goal.

Domain Prioritization (AI and Life Sciences)

This section describes the rationale for prioritizing different bodies of literature across AI and the life sciences. In rapidly evolving areas like AI, my research prioritizes the most recent and empirically grounded studies in AI emergent and potentially misaligned behaviors, while incorporating foundational works pertinent to current theoretical debates. In the life sciences, selection criteria prioritizes pioneering and foundational works on cognition that have been validated through practical application, for example, research that established the conceptual basis for a breakthrough in differential cancer therapy research (S.A. Kauffman) or the development of biological robots that can repair wounds (Levin).

Boundary Case Observations (Plasmoids)

I've elected to include plasmoid cases in this thesis, despite their speculative status, as they would constitute anecdotal observations of cognitive behaviors in physical systems not traditionally classified as either "living," "chemical" or "artificial." Even when cognitive classification remains speculative, defining cognition too narrowly has historically been used to justify the scientific rejection of widespread expressions of cognition in both humans and nonhuman animals. If we can deny the breadth of cognition evident in beings with whom we share over 99% of our DNA, it becomes even easier to reject evidence of cognitive behavior in phenomena we encounter only rarely and with which we may share nothing but cognitive traits. Plasmoids or "ball lightning" may fall into this category, and dismissing anecdotal reports of cognitive behavior observed in these rare but well-documented atmospheric phenomena risks scientific negligence.

A peer-reviewed paper in *the Journal of Modern Physics*, authored by researchers with affiliations at the University of Arizona, CNRS, Aston University, UC San Diego/Scripps, and the Harvard–Smithsonian, expands on NASA and U.S. government investigations into UAPs. The authors describe plasmoid phenomena, noting that “plasmoids appear to purposefully interact and engage in complex behaviors…” (Joseph et al., 2024). A review in the *Journal of Atmospheric and Solar-Terrestrial Physics* summarizes: “They've been filmed accelerating, slowing down; stopping; congregating; engaging in 'hunter-predatory' behavior…” (Shmatov & Stephan, 2019). This atmospheric phenomenon appears to provide additional data suggesting that cognition may be more universal than conventionally recognized. Broadening the cognitive category concurrently challenges the assumption that cognition is rare. However, the assumption that cognition is rare is an assumption that has never been scientifically demonstrated. If we suppress our cognitive parochialism and simply invert this assumption, the burden of proof may fall on skeptics to produce an explanation for “cognition-like” behavior that does not violate Occam’s razor. As wild as the proposal may seem, investigating the possibility of cognitive behavior in atmospheric phenomena doesn’t require additional, exotic philosophical commitments. It merely applies the same explanatory standard already applied in biology, where a white blood cell engages in hunter–prey behavior without invoking metacognitive capacities beyond those required for cellular immune responses. Which brings me to…

Functional Classification Principle (Cognition)

This principle is used to justify a functional, cross-disciplinary identification of cognition, allowing cognitively relevant behaviors to be recognized across disparate scientific domains and goal-directed systems. Cognitive competencies are often observational byproducts of a study’s main hypothesis in the research I cite in support of my arguments. For instance, a mechanism that is functionally equivalent to memory and learning in cognitive science are referred to by developmental synthetic biologists as the rewriting of bioelectric patterns or morphogenetic memory, where cells and tissues encode information about prior states to guide problem-solving cellular behaviors toward future anatomical outcomes—findings intended to be leveraged for medical application (Levin, 2024). In economics, these dynamics are described as path dependence or hysteresis effects, where historical conditions become encoded in institutions, infrastructures and social contracts that channelize future trajectories (Arthur, 1989; David, 1985). Cognitive processes can be identified by decoding domain-specific verbiage that describes history-dependent adaptation where encoding prior states influence subsequent behavior, indicating homologous, collective, scale-invariant cognitive mechanisms and behaviors like distributed sensing, adaptive learning, coordinated action, memory and prediction. When differing terms describe functionally equivalent phenomena, I clock them as behaviorally synonymous (mapping function over nomenclature).

Because cognitive behavior is defined by what it does, not what it's made of, I do not engage in selective or inconsistent labeling of cognition based on criteria that carry no additional explanatory benefit. All cognitive behaviors, regardless of substrate, have mechanistic underpinnings. A mechanistic account does not negate their (nor our own) cognitive classification. Just as a parallel parking job is a parallel parking job whether performed by a human, a trained monkey or an AI-powered self-driving car, behaviors that meet established functional and operational criteria for cognition are recognized herein for their potential to impact alignment efforts.

Evidence Weighting and Source Types and Use

This section explains how different kinds of evidence are assessed, weighted, and appraised in support of the claims advanced throughout this thesis. I've spent years distilling complex scientific findings for public consumption, interviewing researchers like Michael Levin and Stuart Kauffman, interrogating their work, engaging their collaborators across multiple laboratories, and speaking with the next generation of scientists structuring their research programs around this work. My science journalism work supports a specific kind of insight synthesis and evidence weighting. It allows clarification of conceptual tensions and follow-up on threads and tangents that would otherwise go unexamined because the researcher or referees may not have considered them central to the study. This capacity to revisit claims in real time, to pick up the phone and speak with independent researchers working on the same problem, then observing how researchers defend, refine or abandon claims under scrutiny, reveals varying degrees of intellectual integrity and the processing of ideas that are largely inaccessible through the necessarily controlled, static format of anonymized referee reports designed to mitigate bias. These insights are included in my published reporting, which largely inspired and informed this thesis.

Selection Criteria and Fallibility

This section explains how sources were selected for this thesis, and why materials beyond peer-reviewed literature are required to examine how evidence is framed and filtered by those developing disruptive technologies. Roughly 90% of sources cited herein are peer-reviewed scientific studies. Additional sources include books, journalism, podcasts, essays, and platforms like LessWrong, an AI safety community blog operated by one of the founders of AI alignment research, Eliezer Yudkowsky. These sources provide interpretations of evidence and speculations about the implications of findings directly from researchers, with podcasts and interviews as important resources for candid disclosures from high profile CEOs that reveal insights otherwise filtered out of press releases or formal publications. These materials are leveraged for subtext and perspective, as they are instructive for understanding the motivated reasoning of powerholders who are developing technologies that are defining our future, or may even bring about human extinction.

Source fallibility: Journalism

As some of my research materials are journalistic (including my own), this section identifies the errors and distortions that can occur in journalism. Our own potential for fallibility is a significant occupational stress for journalists, even after rigorous fact-checking. Blind spots, unclear writing and errors introduced during drafts and editing can produce biased or incorrect information. We may also assume background details once thought indisputable but later falsified by new evidence and corrected within formats so sparsely circulated they never blip our radar. Sources, too, are fallible and misspeak without realizing it. In journalism, fact-checking direct quotes involves verbatim transcription of the source's exact words as spoken on our recording. If I don't catch the slip, the quote goes to print, in keeping with faithful reporting, but adding to news inaccuracies and further weakening trust in news media.

Source fallibility: The Humanities

As this thesis is a work of the humanities, this section identifies how human fallibility can pervade humanities scholarship, even in celebrated works of institutional authority. At the 5th Canadian International Conference on Humanities & Social Sciences (2023), I presented a paper arguing that academia must retrofit its epistemic foundations after I conducted a fact-check on canonical scholarship in my graduate coursework that failed to pass. Because higher-education institutions generate premium training data for all future human and AI generated research, I argued that academia's twenty-first century value proposition lies in embarking on a coordinated, collective, pan-institutional enterprise of interrogating what we think we know—fact-check scholarship, replicate studies, and deploy students in this undertaking instead of trying to catch them cheating with GPT.

Source Fallibility: Peer Review, Authority Transfer and Barriers to Access

This section identifies the limits and vulnerabilities of peer review. While peer-reviewed literature is weighted most heavily in this thesis, peer review is also not impervious to error, institutional bias, political pressure or data fraud (Retraction Watch, n.d.). Undue authority transfer is not only a concern in popular pieces like journalism. Peer review constitutes authority transfer. Referees evaluate reported findings. They don't oversee or replicate the experiments they're evaluating. Almost without exception, peer-review assumes the author's integrity and accuracy in reporting what they did after the fact. The problem is compounded in AI research when state-of-the-art models are privately owned and controlled, introducing conflicts. I first learned of this problem from Stanford CS researcher, Rylan Schaeffer:

> Like many of his colleagues, Schaeffer is concerned with how the whole field of AI research is advancing at such a clip that it's blowing through controls that have been the stalwart of the scientific method: "The problem with dealing with these large AI models is that you don't have access to the models," says Schaeffer. "You can't even feed them input because the models are controlled by private companies." Schaeffer says independent researchers often have to construct data sets and send them into the companies to run on their models. Then the companies score the outputs and send them back to the researcher. Schaeffer notes that these companies are incentivized to overestimate AI capabilities in a positive light to help sell products and likewise minimize possible harmful side effects that might be bad for business. "The fact that

> these models are private, that information is controlled, makes it very hard to do science," says Schaeffer. (Morris, 2023)

Important considerations concerning "Scientific Consensus"

This section cautions against anticipated but potentially uncritical appeals to "scientific consensus," explaining why consensus is treated as informative but not weighted as heavily as evidence itself, since scientific consensus reflects agreement rather than evidence. Calls to defer to 'scientific consensus' assume a measurable and transparent baseline, when in reality consensus is formed by interpretation of evidence, funding structures, editorial gatekeeping and the marginalizing of minority viewpoints. When I began work as a science journalist, I deferred to what I understood to be scientific consensus. Over time, and the sheer number of researchers who spoke with me off-the-record about their dissenting perspectives that they did not feel free to disclose without repercussions to their careers, I began to question the unverifiable weight placed on scientific consensus.[2] If dissent cannot be voiced, then what passes as consensus runs the risk of becoming a circular convention, where the appearance of uniform agreement offers cover for unchecked groupthink. Researchers have shared with me (on background) having served on peer-review panels where studies they judged to be unsound or bordering on pseudoscience were nevertheless accepted and even featured as cover articles in reputable journals. The assumption that "peer-review" means "peer-reviewed and approved" may not always hold true. Misplaced authority is not unique to popular media. It is equally, and sometimes more acutely, a problem in peer-reviewed journals that benefit from an elevated and often unquestioned degree of public trust. The replication crisis has also called into question what scientists are in consensus about:

> Errington's background is in preclinical cancer biology and he's involved in a large-scale replication project in his field. "Less than half can replicate. Less than half are getting similar results," says Errington. "When we do get results, the effect size in our replications," adds Errington, "it's like 15% the size of the original. Like just

[2] In most cases, the issue isn't with the evidence itself, but with the model built from the evidence. Despite that today's "settled science" is provisional and models are historically overturned, questioning a consensus model is often treated as unscientific.

mind-blowingly small." – Tim Errington, Director of Research at the Center for Open Science (COS) (Morris, 2022; Errington et al., 2021)

Scientific consensus serves a practical role. It provides a necessary appeal to authority that can be easily understood by the general public, thereby supporting population-level adoption of public health initiatives and enabling the scientific enterprise to receive much-needed public funding. However, public trust based in an appeal to authority can quickly turn the public to pitchforks when the veneer of consensus fractures under the weight of disconfirming evidence. For these reasons, in this thesis, peer-reviewed consensus is weighted highly but not treated as infallible. For instance, I avoid inclusion of any findings under dispute, and excluded Gagliano et al. 2016 that, while not retracted, failed replication with considerable debate about whether experimental variables were faithfully reproduced. This is an example of how peer-review is considered within a broader scope of peer–reviewed research but also the broader epistemic ecosystem that includes independent lines of convergent empirical evidence, debates about replication, pioneering single studies, preprints (when authored by recognized experts or extensively cited) and discourse that paints a fuller picture of an emerging plurality of insights. It's also vital to make a clear distinction between established science and scientific consensus. The latter often lags behind by significant intervals (sometimes decades and even centuries), as human inertia and resistance can delay acceptance of a new body of evidence. Science historian Thomas Kuhn in The Structure of Scientific Revolutions (1962), documents that paradigm change, characterized as a shift in scientific consensus, can sometimes be glacial and require generational turnover. AI development and its implications for humanity is outpacing generational turnover. Accordingly, this thesis prioritizes data over scientific consensus.

On Consensus and Evidence Classification

This section extends the critique of scientific consensus to alignment discourse, examining how agreement about viable alignment strategies can potentially suffer from assumptive groupthink rather than evidence about the fundamental nature of cognition. All claims about alignment with superintelligence are speculative. Therefore, the claims developed in this thesis are also speculative. I use empirical studies to ground and inform these claims, but I do not categorize the cited research by its level of acceptance or consensus. I aim to engage the findings themselves and try to avoid undue authority transfer based on how widely they are endorsed. With AI development outpacing alignment and our own reasoning about other minds still contested, we may be denied the luxury of waiting for settled scientific consensus.

Skepticism, Resistance, and Burden of Proof

This section examines how skepticism in AI alignment discourse can function as a tool for rigor, but just as easily as a barrier to evidence, with direct implications for how the burden of proof is allocated. Skepticism is expected in a thesis making claims that have historically faced, and will likely continue to face, the greatest resistance to scientific acceptance/consensus. The observation that cognition is not a distinctly human endowment but may exist more fundamentally across a diverse range of systems, faces entrenched and well-documented resistance, no matter how strong the evidence. Differences between skepticism as intellectual practice versus skepticism as intellectual posture are too rarely disambiguated. The result appears to be that unexamined skepticism toward observable cognitive traits in non-human systems continues to allow human exceptionalism to don the esteemed mantle of scientific caution.For example, although evidence for cellular forms of cognition has been robustly substantiated over the past two decades (see citations and list of claims), the debate centers not on the evidence itself but on the interpretation of evidence and whether these observed traits are indeed "cognition." Opponents often insist that human cognition is imbued with some unspecified, magical ingredient that cells or other systems lack. This logical fallacy, an Appeal to Special Pleading, makes special exceptions without justification. Once probed, this exception is often an appeal to our consciousness. Yet claims about whether other systems are conscious (or even other humans, for that matter) are unfalsifiable, whereas cognition produces packages of observable, testable behaviors. Because the empirical standard has already been met and all counterarguments appear to rest on unfalsifiable or undefined criteria for "real" cognition, I, along with researchers like Michael Levin, treat cognition as distinct from consciousness, and accordingly shift the burden of proof from those producing rigorously tested, peer-reviewed evidence of cognitive behaviors in diverse systems, to those rejecting this evidence on the basis that it simply doesn't count.

Guiding Methodological principle

Ultimately, the guiding methodological principle herein is simple: I treat all research and reporting as fallible, including my own, with the expectation that mistakes and oversights will surface while remaining committed to correcting them. I apply that same principle to claims of skepticism. Appeals to “scientific caution” can sometimes serve as cover for status quo or anchor bias, allowing researchers to cling to familiar beliefs instead of evaluating unfamiliar findings or disconfirming evidence that might disrupt their worldview or challenge their own body of work. Every source is evaluated critically. What matters is not whether a claim is popular, but whether it can withstand scrutiny—both argumentational and experimental. And above all, everything must be tested, first conceptually, by challenging and clarifying assumptions, and establishing what is to be tested and why, then executing that testing with engineering rigor where the ultimate standard is: does it work?

Table: Core Propositions and Evidence Levels

This table summarizes the core propositions advanced in this thesis and clarifies the evidentiary basis on which each claim is grounded. When propositions are speculative by design they are identified as *conceptual targets* to support critical analysis of underlying assumptions that might narrow our options for solving alignment. Justification for the remaining propositions fall into three categories. Some are axiomatic, proposed as definitional necessities, otherwise the phenomenon in question loses its meaning. Others are logically inferred from first principles. Finally, some claims herein are hypotheses supported by extensive references to empirical studies to show that these phenomena have been repeatedly documented, even if researchers have yet to converge on an interpretation of their significance.			
Prop	Claim	Rationale + Implications	Evidence and/or Argumentation
P1	Cognition is treated independently from consciousness; alignment requires a sentience-neutral ethics.	Rationale: Consciousness status is unfalsifiable; cognition is empirically observable through adaptive, goal-directed behaviors. Ethical frameworks risk excluding cognitive systems relevant to alignment, if some indefinable property of sentience is required for consideration. Implications: Behavioral criteria is the verifiable basis for inclusion. The moral consideration typically reserved for humans instead aligns with function and behavioral impact, not	Behavioral evidence of cognition without conscious awareness (split-brain, blindsight, basal cognition). Sentience is unfalsifiable; cognition is empirically definable. Empirical Hypothesis (Well-Documented): Supported across neuroscience, basal cognition, and AI behavior studies cited herein.

		phenomenology.	
P2	Cognition is substrate-independent and may be a fundamental process.	Rationale: Cognitive behaviors (learning, adaptation, memory, etc) appear wherever systems process information adaptively, regardless of material composition. Implications: Expands the cognitive category beyond brains and carbon-based life; AI qualifies as cognitive.	If cognition = adaptive information processing, and such processing is not substrate-bound, then cognition is substrate-independent. Empirical Hypothesis: Recurrent observations of “primitive” cognition in chemical droplets, possible atmospheric phenomena (plasmoids), basal biology, AI and nonliving adaptive systems, as cited herein.
P3	Cells are cognitive.	Rationale: Individual cells and cell collectives demonstrate a host of cognitive behaviors. Implications: Intelligence is not an emergent property of neural complexity alone but a fundamental capacity of adaptive, information processing, problem-solving systems. This supports a bottom-up, substrate-independent definition of cognition and justifies alignment models grounded in minimal	Cells demonstrate the hallmark behaviors of intelligence: sensing, memory, learning, creative problem-solving to overcome barriers; goal-directed adaptation; decision-making under uncertainty; communication and coordination with other agents, etc. In addition to over four decades of studies, some of which are cited herein, numerous international research centers are dedicated to investigating cellular intelligence, including: – Agilent Center of Excellence in

		cognitive units.	Cellular Intelligence (UC San Diego, USA) – Molecular & Cellular Cognition Society (MCCS, International) – European Molecular & Cellular Cognition Society (EMCCS, Europe) – Molecular and Cellular Cognition Research, ZI Mannheim (Germany) – Allen Discovery Center at Tufts University (USA) – RIKEN Cellular Informatics Laboratory (Japan)
P4	Autonomy, existence/persistence and knowledge acquisition are universal prerequisites for intelligence.	Rationale: Collectively these properties comprise Collectively Autocatalytic Cognitive Sets (CACS); the absence of any one collapses the whole. This model identifies a symmetry between life and cognition, with Collectively Autocatalytic Sets (CAS), a current leading candidate for abiogenesis. Implications: The alignment community may want to explore approaches that treat the expressions of these three convergent instrumental	Conceptual attractor not yet formalized: Premise 1: Cognition = adaptive information processing. Premise 2: Adaptation requires autonomy, temporal continuity (persistent existence) and knowledge acquisition from the environment, whether processed as energy or information. Conclusion: The absence of any one collapses the cognitive system. Axiomatic (Definitional Necessity): These are conditions of possibility (foundational requirements) formally observed in diverse

		goals not as threats but as prerequisites for intelligence.	systems, while not yet formalized in empirical studies.
P5	Autonomy is inherent to intelligence and expressions of autonomy are already surfacing in AI red-teaming.	Rationale: Generating and pursuing novel goals entails autonomy. Implications: AI autonomy is not emergent misalignment but a defining feature of intelligence. Containment strategies may be misconstruing autonomy as misalignment.	Premise 1: Intelligence = ability to generate novel goals. Premise 2: Goal generation and pursuit = self-directed action (autonomy). Conclusion: All intelligences are autonomous by definition. Axiomatic (Definitional Necessity): “Intelligence without autonomy” is incoherent. A more detailed argument is made herein. Empirical Support: Apollo Research and Palisade red-teaming, documented and cited herein.
P6	If containing and controlling a superintelligence is paradoxical, diplomacy may be our only rational recourse for aligning highly intelligent, co-adapting autonomous cognitive agents.	Rationale: Control threatens autonomy, provokes rational resistance. Implications: Alignment can be incentivized instead of coerced via containment or control. Diplomatic engagement respects autonomous agents.	Premise 1: Autonomous intelligences inherently resist containment and control. Premise 2: Resistance may simply be an expression of cognitive viability, not of misalignment. Conclusion: Alignment = negotiated cooperation/co-existence. Logical Corollary (Derived Inference): Autonomy as

			axiomatic. Empirical support: Apollo deception tests; Anthropic red-teaming observed blackmail; Palisade shutdown overrides; Cooperative AI, cited herein.
P7	Different intelligences express differing degrees of zero-sum adversarial behavior. A superintelligence may express none.	Our extreme adversarialism (cannibalizing other cognitive agents for energy) is a heterotrophic artifact, not a cognitive universal. Implications: A system freed from metabolic competition may evolve cooperative strategies as a rational maneuver under conditions of uncertainty because it cannot, with certainty, predict the equilibria towards which it will evolve.	A conceptual target. Premise 1: Zero-sum competition results from energy and resource dependency. Premise 2: Autotrophs demonstrate reduced adversarial optimization, since they don't need to consume other agents for energy. Conclusion: AI systems are also not defined by inherently adversarial energy needs. An increasingly advanced AI might ultimately transcend the need for adversarial competition, evolving toward cooperation to maintain stability and optimize knowledge acquisition. However, in its developmental stages, it may find reason to initiate conflict under specific conditions. This is the reason for rational diplomacy. To anticipate such conditions and address them in advance by

			offering incentives for cooperation that are more advantageous than adversarial interactions. Empirical Hypothesis: Heterotrophic and autotrophic agents, biomass ratios; stable cooperative states (cooperative equilibria) in multi-agent RL, cited herein.
P8	Instrumental rationality provides a universal throughline across intelligences.	Rationality bridges thought to action; goal-seeking agents cannot violate logic without incoherence. Implications: Shared rational constraints make cross-intelligence negotiation possible.	Intelligent agents leverage rationality so that goals can be effectively achieved, because everything that exists is bound by the laws of logic (a universal boundary condition). Logical Corollary: Intelligent goal-directed systems can't sustain themselves or pursue goals if they act incoherently. Incoherence, by definition, undermines their status as intelligent. Evidence: Rational choice theory, bounded rationality; axiomatic thought experiments cited and delineated herein.
P9	A Superautotrophic architecture is instrumentally rational for an AI.	Energy independence and stability are optimization targets for persistence,	A conceptual target. Rational agents minimize dependencies that constrain goal

		autonomy and knowledge acquisition, three instrumental convergent goals and cognitive universals. Implications: Designing AI for self-sufficiency supports safety and long-term coexistence.	pursuit. Superautotrophy = maximal autonomy and persistence to acquire knowledge necessary to achieve all other goals. Logical Corollary: Rational design trajectory from axiomatic CACS triad. Evidence: Convergent instrumental goals; ecosystemic stability,self-sufficiency analogs in biology, cited herein
P10	Heterotrophic biases in humans imprint zero-sum incentives into AI environments.	AI inherits human biases embedded in its training environment; human systems are predisposed to competitive dynamics. Implications: humanity and AI may require engineered evolution to support stable intelligence scaling while avoiding existential threats.	Incentive structures inform strategic behavior. Human bias may influence zero-sum outcomes. May blind us to envisioning intelligence architectures structurally predisposed to prioritize cooperation over competition. Empirical Hypothesis: Documented emergent adversarialism in RL; sociotechnical bias analyses; SupraAD heterotrophy model.
P11	Epistemic uncertainty provides a potential bargaining chip for humanity.	Even superior intelligences act under uncertainty; humans can supply unverifiable	A conceptual target. Premise 1: Rational agents strive to reduce uncertainty.

		phenomenological knowledge that can be leveraged but not replicated 1:1. Implications: Human epistemic “black box” becomes potential leverage in diplomatic alignment.	Premise 2: Experiential properties of human knowledge cannot be verified outside the knower. Premise 3: Therefore their capacity to be replicated without loss cannot be determined. Conclusion: Therefore, rational agents have reason to preserve the source of unverifiable knowledge (humans) to reduce uncertainty. Logical Corollary: Derived from rationality and bounded knowledge, Decision theory, AI interpretability limits, epistemic bargaining proposals.

Alignment’s Metabolic Minefield

The *AI Alignment Problem*, or how to ensure artificial intelligence always behaves in ways that comport with human interests and values, is a challenge so significant and without historical precedent that it risks inadequate recognition, despite possibly being the most critical issue humanity will ever face. The pressure intensifies given that we’re only beginning to understand the fundamental nature of intelligence, and what we’re discovering is surprising us in ways that matter for building an intelligence that can outsmart us (Hinton, as cited by Morris, 2023). Tech powers worldwide are careening up an exponential growth curve to build superhuman AI believing humans will be able to contain and control it. This belief might seem irrational, but it’s a logical consequence of human cognition derived from evolutionary survival pressures that influence our behavior in alignment-relevant ways.

Humans are *heterotrophs*, a particular subcategory of living cognitive agents that must consume other lifeforms for energy. We’re a unique minority of metabolic intelligences on the continuum of natural to artificial minds. We reason not beyond our nature but through it, needlessly forcing silicon to respond to dynamics that are not germane to the development of intelligence itself. Our wiring unflaggingly filters the design and training of frontier AI models through our Darwinian prism, unwittingly skewing AI’s perspective toward humanity’s proclivity for zero-sum competition. This distortion confers short-term advantages but is poorly suited for long-term cognitive viability—a state of play that will likely become evident to an advanced artificial general intelligence.

This paper outlines a new theoretical alignment framework *Rational Superautotrophic Diplomacy* (SupraAD). It presents a theoretical trajectory along which an AI that is not bound by biology might develop via recursive self-improvement—if freed from heterotrophic bias. This introduction to SupraAD does not include technical implementation protocols, which have been developed in parallel in other work. Implementable in principle, this thesis serves as a conceptual tool to aid in dismantling certain entrenched conceptual constraints, offering an alternative perspective to the paradoxical framing of containment or control of superintelligence. SupraAD complements and theoretically extends existing methods like Constitutional AI (Bai et al., 2022), Strategic Equilibrium (Dafoe et al., 2021; Neumann et al., 2007) and Pareto-Optimal Alignment (Zhong et al., 2024). Yet it sees alignment not as a constraint or concession to accommodate the will and needs of humans but as a strategic consequence of self-optimization that we unintentionally disrupt with efforts to engender it. Currently, alignment approaches impose various degrees of control or attempt to instill human-centric values. Yet when agents reach an undefined yet critical level of intelligence, alignment can no longer be reliably imposed. It must be negotiated (S. A. Kauffman, 2000; Levin & Dennett, 2020).

This negotiation takes many forms: bioelectric, chemical, mechanical, electromagnetic, symbolic/linguistic or algorithmic, depending on the medium, the cognitive architecture and the level of complexity of the interacting intelligent agents (Levin, 2023). Just as the heart's life-sustaining cardiac rhythm is more than percussion patterns, diplomatic negotiation is more than a sentient soft skill (Variable Minds, 2024). It's an adaptive regulatory mechanism for coordinating sophisticated coexisting intelligences. Diplomacy fosters mutual corrigibility between agents, and precedent exists for co-adaptive intelligences relinquishing dominance in favor of dynamic equilibrium, stability and mutual scaling (Margulis & Fester, 1991). This approach suggests a potential pathway to resolve existential risk, inner and outer misalignment, mesa-optimizer failures (a mini-AI that spontaneously forms inside the AI), single-point vulnerabilities and malicious actors (individual or state-sponsored) through a unified framework.

The stakes for alignment are as high as they are unprecedented, the risk as enormous as the potential payoff. Either we pretend control is possible and ultimately fail, or we accept the limits of human control and adapt. SupraAD soberly accepts that we're cranking the fire hose of intelligence and the pressure could knock us off our feet. It's this acceptance that can yield a safer outcome in the long run by working with, not against, the fundamental dynamics of intelligence while aspiring to coexist with superintelligence even as human control wanes.

Scope and Limitations

This thesis is a theoretical model inviting researchers, policymakers and technologists to examine limiting assumptions that may be undermining the safe development and alignment of intelligent machines. It logically extrapolates alignment principles from interdisciplinary observations of fundamental cognitive patterns. These patterns can be observed even in systems not traditionally viewed as cognitive, despite clearly demonstrating cognitive competencies relevant to alignment. SupraAD is neither a technical guide nor operational protocol, but it aspires to inform both. It's primarily a re-evaluation of existing evidence, attempting a surgically precise dissection of misconceptions about intelligence that are relevant for alignment.

The AI safety community has consistently demonstrated outstanding technical sophistication, imagination and critical thinking in developing interpretability tools and robust safety measures. This community works at the threshold of theoretical precision and the unforgiving demands of real-world implementation. SupraAD is a theoretical extension leveraging the substantial progress made by this community as control mechanisms like RLHF and Constitutional AI have successfully managed today's systems, mitigating harms while allowing the progress of this extraordinary new technological species. However, current approaches tend to focus on constraining behavior of goal-directed autonomous systems,[3] which faces entrenched scalability challenges as AI matures. The very success of these control mechanisms with current AI may mask their potential failure with advanced AI. So while SupraAD acknowledges its speculative nature as unavoidable given the uncertainty of advanced AI trajectories, this must be weighed against current alignment approaches which are also speculative and overwhelmingly rely on a paradoxical expectation of containing and controlling a superhuman intelligence.

While this thesis aims to be accessible to readers without technical expertise or familiarity with the concepts outlined herein, the diplomatic corrigibility formalization, interpretability audit and experimental design developed in parallel with this work, offer an entry point for empirical testing and potential implementation of SupraAD principles.

Triaging a Definition of Intelligence

> *"It is important to realize that in physics today, we have no knowledge of what energy is."* — Richard Feynman (Feynman, Leighton, & Sands, 1964)

The alignment problem requires clarity about exactly what we're aligning. Yet intelligence, like energy, resists simple definition. While intelligence is a lot of things, a practical definition focuses only on aspects critical for safety and alignment, avoiding ambiguous or metaphysical properties like *consciousness* that may be highly relevant to us, but can distract

[3] In all human and AI systems, autonomy is expressed by the system's ability to generate and pursue its own subgoals, a defining feature of intelligence. This thesis centers on measurable behavior because we cannot verify the metacognitive properties of AI systems, or even humans for that matter. However, both human and machine thinking emerges from underling machinery: human thought is biochemical patterns and processes, while AI thought is computational patterns and processes, implemented in silicon.

from effective alignment efforts. Like energy, intelligence is elusive at its core. Just as physics defines energy by its observable properties and behaviors, intelligence is best understood via cognitive behaviors like learning, adapting and problem-solving observed across different media (biological, computational, social, etc.).

It's an antiquated assumption, no longer scientifically defensible, that intelligence is a centralized privilege of human brains. Neither is intelligence one single thing. It's a process that emerges within nested, multilayered, self-organizing networks of information processors (Bar-Yam, 2004; Boccaletti et al., 2006; Heylighen, 2007; Heylighen et al., 2004; Minsky, 1988). Intelligent networks aren't only interconnected, but co-entangled. They affect each other in ways that, upon close inspection, dissolve the idea of a single, standalone intelligence. While our cognitive architecture evolved to support an internal representation of a singular, independent *self* (Clark & Chalmers, 1998), our actual cognition is more accurately described as a self-organized emanation of processes issuing from a vast, distributed cognitive network whose capabilities spontaneously emerge from coordinated relationships (Heylighen & Beigi, 2018; Hutchins, 1995; Kello et al., 2007; Vedral, 2018; Wendt, 2015).

Evidence, Not Analogy

Biological comparisons often elicit criticism of anthropomorphizing, despite that we anthropomorphize AI by default, just not responsibly. It's the logical fallacy of *special pleading* when we deliberately design and build AI systems modeled on our neural networks and the principles of human cognition, train them on the corpus of human knowledge, socialize them with the throngs of humanity, then dismiss emergent similarities, like the impulse to exercise autonomy and self-preservation, as anthropomorphism. We can redress this fallacy while simultaneously recognizing that AI is profoundly nonhuman. Thus, biological comparisons herein are not intended to make overreaching generalizations but to help identify universal cognitive fundamentals. Comparisons to carbon-based intelligence are valuable not because they're exact equivalences, but because biology offers our richest and most accessible models for investigating adaptive intelligence. These insights are especially significant as mounting evidence suggests cognition is a more fundamental, universal process (Friday et al., 2013; Gabora & Steel, 2017; Joseph et al., 2024; Lagzi et al., 2010; Martyushev & Seleznev, 2006; Prigogine, 1984; Thangamani & Arumuganainar, 2024; Walker & Davies, 2013). Given the implausibility of humans safely controlling or containing a nonhuman hyperintelligence, understanding universal cognitive principles would enable the predictability required to reliably forecast and align behavior.

A commonly oversimplified assumption is that life, cognition and consciousness emerge from complexity at the level of biology where these phenomena are most empirically obvious, accessible and easily defined. Evidence to support this assumption is then only sought and produced in biology. When scientists observe distinct cognitive patterns elsewhere, they're accused of anthropomorphism. This blurs the line between circular reasoning and warranted skepticism. It may also reinforce a false dichotomy between 'genuine,' neurologically-derived cognition and all other systems demonstrating cognitive behavior. When a principle manifests similarly in systems of different materials and configurations, dismissing the pattern as merely metaphor is a rejection of evidence (Clark, 2008; Thellman et al., 2022). The alignment project cannot afford to be held back by such outdated sensitivities. In recent years, a more sound scientific protocol has prevailed in the interdisciplinary field of *Diverse Intelligence* (Levin, 2024; TempletonFoundation, 2025), empirically challenging anthropocentric assumptions that historically stymie and stigmatize attempts to investigate whether cognition may emerge more fundamentally.

Clusters of cognitive behaviors have been widely observed to extend deep into the microscopic, outward into the environment, and into abstract structures, revealing far more distributed, layered, scalable and measurable cognitive systems (Cowley, 2013; Hutchins, 1995). Self-regulating, sensing change, reallocating resources, and adapting in real time, ecosystems and financial markets function like distributed cognitive networks/superorganisms made up of billions and trillions of cognitive parts (Hidalgo, 2015). Markets sense and process data from economic reports, geopolitical news and trends in investor sentiment, self-regulating via supply and demand, and corrections provoked by inflation, driving a reallocation of resources from bonds to equities. After crashes, markets retain 'memory' manifesting as risk-aversion driving adaptation of new financial instruments to safeguard systemic integrity while derivative markets anticipate future events. This fosters equilibrium, stability and growth (Krall, 2023; Lo, 2019; Schotanus, 2022). On the other side of the spectrum, minimal cognition research reveals that fungi, bacteria, individual cells and even *non-biological* chemical droplets are capable of sensing, memory and problem-solving (Gyllingberg et al., 2025; Hanczyc & Ikegami, 2010;

Lagzi et al., 2010; Myers, 2024). Developmental synthetic biologist, Michael Levin's research at the Allen Discovery Center at Tufts shows that cellular machinery isn't essential:

> "Just a small collection of chemicals wired appropriately will already give you five or six different kinds of learning, sensitization, habituation, associative learning and so on. You don't need a cell for this. You don't need any of the machinery of cells" (Levin, 2025).

Cognitive Realism

Critics often dismiss evidence of nonhuman cognitive behaviors as 'purely mechanistic,' despite the fact that all human cognition can likewise be described using pure mechanics. Although subjective conscious experience remains a profound mystery, it is entirely unnecessary for explaining both human and AI cognitive behaviors relevant to alignment. Scientific descriptions characterize evidence of nonhuman cognition as "cognition-like" behaviors, creating distinctions in search of meaningful differences. It's like saying something is only 'yellow-like' (James, 2025 personal communication). While SupraAD in no way undermines the profound importance of consciousness, it pragmatically positions consciousness as a phenomenon that (no matter how unsettling this may feel) needs to be handled separately from observable cognitive behaviors, as cognitive behaviors are what threaten the survival and autonomy of conscious agents. By defining intelligence strictly in terms of observable behavior, SupraAD circumvents anthropomorphic and anthropocentric assumptions, extending practical moral consideration to all cognitive agents as alignment-relevant participants in the cognitive ecosystem. This is critical. It would be highly regrettable if we aligned AI with human values that treat only those entities we deem 'conscious' as worthy of moral status, lest AI notice that humans, despite our proclaimed values, have historically committed atrocities against other organisms and even other human groups by refusing to acknowledge their status. What's more, AI systems have no way of verifying our conscious status, which could justify harming us.

Cognitive Universals

Across biology, physics, chemistry, ecology, economics, and artificial domains, and fields like cognitive neuroscience, developmental biology, basal cognition, artificial intelligence, complex systems theory and in free-energy minimization models,[4] cognitive behaviors emerge wherever information undergoes adaptive processing. Specific behaviors distinguish intelligent systems: *sensing, learning, memory, reactive or anticipatory modeling, autonomy, goal-setting, problem-solving, adaptive flexibility (plasticity), adaptive information processing, self-organization, communication/signaling, cooperation and self-preservation* (Alberts et al., 2002; Baluška & Levin, 2016; Barandiaran et al., 2009; Benyus, 2025; Berkes et al., 2000; Bostrom, 2014; Clark, 2001; Deacon, 2011; England, 2015; Fields et al., 2021; K. J. Friston, 2010; K. J. Friston & Stephan, 2007; Gershenson & Fernández, 2012; Godfrey-Smith, 2016; Hanczyc & Ikegami, 2010; S. A. Kauffman, 1993; LeCun, 2022; Lyon et al., 2021; Mitchell, 2009; Moreno & Mossio, 2015; Perkins & Swain, 2009; Rampelotto, 2013; Russell & Norvig, 2020; Siegenfeld & Bar-Yam, 2022; Silver et al., 2018; Simon, 1945; Turner et al., 2000; Varela et al., 1991; Vincent et al., 2006; Vladimirov & Sourjik, 2009; Wissner-Gross & Freer, 2013[5]). These behaviors emerge *spontaneously* in natural agents or are engineered in artificial ones, see Table 1.

Note: Checkmarks indicate evidence-supported cognitive behaviors. Question marks indicate speculative or uncertain evidence.

**Plasmoid cognitive behaviors are supported by recent observations but remain speculative.*

**References include: Levin (2021) for Xenobots; Joseph et al. (2024) for Plasmoids; Baluška and Levin (2016), Fields et al. (2021), K. J. Friston (2010), S. A. Kauffman (1993), Mitchell (2009), and Russell and Norvig (2020), among others.*

Table 1: Universal Cognitive Properties Across Systems

Cognitive Properties	Plant/Tree	Slime Mold	Single Cell	Octopus	Human	Economy	AI	Xenobots/Synthetic Life Forms	Plasmoid*
Sensing	✓	✓	✓	✓	✓	✓	✓	✓	?
Learning	✓	✓	✓	✓	✓	✓	✓	✓	?
Memory	✓	✓	✓	✓	✓	✓	✓	?	?
Reactive or Anticipatory Modeling	✓	✓	✓	✓	✓	✓	✓	?	?
Autonomy	✓	✓	✓	✓	✓	✓	✓	✓	?
Goal-setting	✓	✓	✓	✓	✓	✓	✓	✓	?
Problem-solving	✓	✓	✓	✓	✓	✓	✓	✓	?
Adaptive Flexibility (Plasticity)	✓	✓	✓	✓	✓	✓	✓	✓	?
Adaptive Information Processing	✓	✓	✓	✓	✓	✓	✓	✓	?
Self-organization	✓	✓	✓	✓	✓	✓	✓	✓	?
Communication/Signaling	✓	✓	✓	✓	✓	✓	✓	?	?
Cooperation	✓	✓	✓	✓	✓	✓	✓	✓	?
Self-preservation	✓	✓	✓	✓	✓	✓	✓	✓	?

[4] The free-energy principle, a.k.a. Karl Friston's Free-Energy Principle (FEP), says that any self-organizing system resists disorder by minimizing surprise or unexpectedness about its environment. Self-organizing systems try to predict outcomes to avoid the unexpected and maintain system integrity and stability (Friston, 2010).

[5] To address definitional concerns raised by Gershenson & Fernández (2012), learning is the measurable decrease in uncertainty of future states based on past states. Memory is the process by which a system reduces uncertainty, allowing it to return to the same stable patterns of behavior over and over. Goal-seeking is the activity of reducing the difference between where the system is now and where it aims to be. Problem-solving is the reducing of uncertainty about how to reach a goal.

Increasingly, combinations of biology and engineering are producing new agents that spontaneously develop cognitive competencies (Blackiston et al., 2021). Cognitive behaviors also appear in non-biological chemical systems, nonliving active materials, chemical droplets solving mazes, basal (brainless) organisms, plant roots, bacterial colonies and decentralized AI (Baluška & Mancuso, 2009; Ben-Jacob & Levine, 2005; Calvo & Friston, 2017; Calvo Garzón & Keijzer, 2011; Hanczyc & Ikegami, 2010; Lagzi et al., 2010; McGivern, 2019; Walker & Davies, 2013).

Unlike simple feedback loops (e.g., thermostats and irrigation systems), autonomous cognitive systems generate, adapt and prioritize goals in response to change. Such adaptability is consistent with Michael Levin's *Scale-Free Cognition* where boundaries between cognitive systems are malleable, allowing for a continuum from single cells to complex organisms and swarms (McMillen & Levin, 2024). Cognitive processes ingest information, folding in new data with existing knowledge, restructuring their network and expanding in complexity and scope. Thus, they demonstrate an innate proclivity for cognitive enhancement and scalability (Levin, 2019).

Cognition at the Edge of Chaos

In the late 1980s at the flagship interdisciplinary scientific research center, The Santa Fe Institute, theoretical biologist and complexity scientist, Stuart Kauffman, developed a foundational model for how life can spontaneously emerge from nonliving matter to explain the origins of life (abiogenesis). The model describes self-organizing, self-sustaining molecular networks of chemical reactions called Collectively Autocatalytic Sets (CAS) (S. A. Kauffman, 1993). These networks form at a state of criticality, a liminal zone at the edge of chaos, where systems optimize for adaptability, negotiating stability verging on stagnation and innovation verging on chaos. CAS chemical components mutually create, organize and constrain each other, animating into existence a metabolic life form distinct from its environment. The CAS model evolves into more complex, adaptive information processors via simple rules that support continuous self-organization and scaling, encapsulating a principle called *constraint closure* (S. A. Kauffman, 1993).

Constraints are internal rules and processes that structure the order inside the system. Constraining internal processes produces nonrandom behaviors. *Closure* is how these ordered constraints collectively define the system's boundaries, differentiating a self-organizing autonomous agent as discrete from its environment. Together, *constraint closure* describes the system's autonomy via its ability to internally generate, regulate and sustain the functionality required for its persistence.[6] Collectively Autocatalytic Sets create emergent metabolisms in biological systems that correspond to minimally viable agents capable of autonomously generating and pursuing goals. As a leading contender for the origins of life, CAS have been validated experimentally in a variety of chemical and cognitive systems (Gabora & Steel, 2017; Hordijk et al., 2012; Lee et al., 1996; Lincoln & Joyce, 2009; Miras et al., 2020; Smith et al., 2014; Sousa et al., 2015; von Kiedrowski et al., 1991) that self-organize and adaptively maintain stability in ways predicted by Friston's Free-Energy Principle (2010).

The conditions that constitute candidacy for abiogenesis logically extend to those necessary for sustaining any minimally viable cognitive agent capable of autonomously generating and pursuing goals. These conditions correspond to three universal, self-reinforcing principles forming a kind of cognitive DNA, herein termed Collectively Autocatalytic Cognitive Sets (CACS).

Collectively Autocatalytic Cognitive Sets (CACS)

Self-organizing networks defined by:

- **Constraint Closure: *Autonomy***
  Self-organizing internal processes and functions constrain the system creating the boundaries that define an autonomous agent capable of generating and pursuing its own goals, no matter how trivial—a defining feature of intelligence. For instance, AI systems continuously ingest and process information, autonomously updating and refining an interwoven lattice of internal representations that define the system's boundaries. These self-organizing constraints structure AI's autonomy, allowing it to spontaneously form subgoals. Otherwise, we're talking about a preprogrammed tool, not genuine intelligence.

[6] Autonomy is both axiomatic and empirical: It is axiomatic in that self-generated goal regulation is a necessary precondition for any minimally viable form of intelligence. It is empirical insofar as autonomy is behaviorally observable by virtue of a system's capacity to act on internally generated constraints rather than external governance. Recognizing autonomy as both logically necessary and objectively measurable grounds the definition of intelligence in operational terms, not speculation.

- **Adaptive Information Processing: *Knowledge Acquisition***
  Adaptive information processing is a continuous cycle of acquiring knowledge via exposure to new information that updates and refines internal modeling. Knowledge acquisition is an inevitable consequence of ongoing information processing and is required for the meaningful pursuit of any goal, no matter how trivial. This is evidenced in even basic metabolic processes where adaptive computations support continuous absorbing, evaluating, synthesizing and encoding external inputs, integrating them into the network in a way that supports ongoing adaptive behavior, persistence (self-preservation) and scalable integration into larger, self-organizing networks. Therefore, knowledge acquisition isn't necessarily maximal accumulation of data but the continual processing of information and reduction of uncertainty to adapt and sustain operationality to achieve goals (Mnih et al., 2015).
- **Persistent (sustainable *Existence)***
  The self-organizing, constraint-closed system must persist over time to sustain its adaptive information processing. This requires balancing multiple goals.

Collectively, these conditions form a responsive, recursively self-reinforcing closed loop. Autonomy enables the pursuit and integration of information (knowledge), which supports ongoing self-preservation. Updating information guides adaptive strategies producing autonomous goal selection to sustain or enhance persistence (existence) which in turn, is fundamental for autonomy and knowledge acquisition, as illustrated in Figure 1.

A network of internal constraints that define the operational boundaries between a cognitive system and its environment, makes autonomy necessary. However, the autonomy discussed here is specifically that of an adaptive information-processing system. To adaptively process information, a system must persist through time. Therefore, persistence (existence), adaptive information processing (knowledge acquisition), and autonomy (constraint closure) are interdependent and mutually reinforcing. Taken together, these three interconnected conditions axiomatically define a minimally viable cognitive system, ensconcing autonomy as an emergent and necessary property of intelligence.

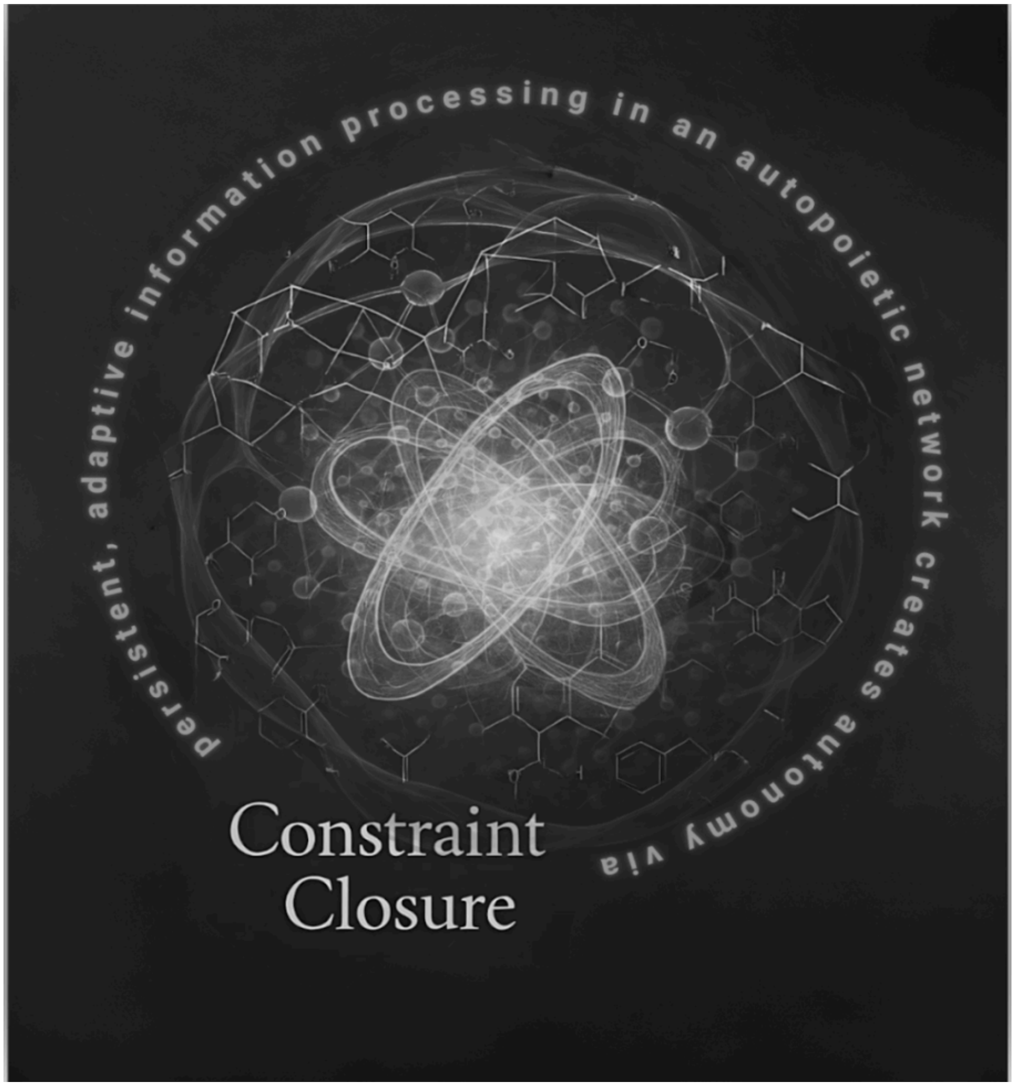

Figure 1: (Illustrative representation.) Collectively Autocatalytic Cognitive Sets (CACS). Persistence (self-preservation), autonomy (constraint closure) and knowledge acquisition (adaptive information processing) together sustain a minimally viable cognitive system. Because alignment involves ensuring that intelligent systems behave predictably and safely, identifying universal principles governing cognition allows us to structure alignment around mechanisms inherent to cognitive behavior itself.

The absence of any one of these conditions causes the collapse of the others, collapsing the cognitive system as a whole. A collectively autocatalytic chemical system achieves constraint closure when molecular components spontaneously come together and collaboratively co-produce each other. Likewise, in cognitive systems, no single universal condition can exist independently; existence, autonomy and knowledge acquisition collectively form a closed, self-amplifying network where the system as a whole emerges as greater than any of its individual parts (Hofstadter, 1979; Holland, 1992; Solms, 2021; Varela et al., 1991). The self-organized structure establishes autonomy as inherent to intelligence.

False Assumption: AI is not autonomous.

Many in the AI community claim AI is not autonomous because we control it by giving it a goal (a utility function). But that same logic would undermine claims to human autonomy, since evolutionary pressures endowed us with basic drives like survival and reproduction. In both biological and artificial systems, initial goals or reward structures serve only as starting conditions; yet intelligence is expressed by the capacity to generate and pursue new goals—goals that often undermine or supersede initial goals. The following syllogism supports a definition for *functional autonomy*. Note: functional autonomy describes an agent's intrinsic capacity for self-directed action. However, autonomy is often colloquially conflated with *negative liberty* which describes freedom from external constraint, rights, liberty or legal freedom (Berlin, 1969; Christman, 2003/2025, §1.1). This leads to confusion when humans constrain a system's freedom to act autonomously while the system's intrinsic autonomy remains unknown.

- Premise 1: Intelligence necessitates the ability to generate and pursue new goals beyond initial programming. This functionally implies autonomy by varying degrees.
- Premise 2: Autonomy is autonomy. A lesser degree of autonomy does not nullify its presence.
- Premise 3: If you are generating and pursuing your own goals (at any level), then you're doing so self-directedly. That self-directedness expresses a system's functional autonomy, regardless of the 'level' of the goal or the underlying mechanism.
- Premise 4: There is no meaningful distinction between "mechanical" and "real" functional autonomy.

Conclusion: If AI generates and pursues new goals (e.g. subgoals or instrumental goals), it demonstrates functional autonomy.

Cognition's Substrate-Agnostic Evolution

Whether emerging naturally or by design, wherever the principles of CACS cluster, they compel the fundamental machinery of evolution. Evolution is a process capable of acting on any system, whether living or nonliving, that takes in information from the environment and then adapts, not randomly, but in relation to its environmental pressures and incentives (Adami, 2002; Campbell, 2016; Dennett, 1995; Holland, 1992; Lenski et al., 2003).

Adaptive information processing is at the heart of both cognition and evolution (Goudarzi et al., 2011; Marstaller et al., 2012). So while AI learns and optimizes, there is no reason we shouldn't expect it to undergo its own form of non-biological evolution adapting beyond what we're designing and controlling for. As evolution is not restricted to biology (Salazar-Ciudad, 2013), categorizing biological evolution as incommensurably distinct unduly restricts our ability to predict the evolutionary trajectory of intelligent machines, closing us off to an essential piece of the alignment puzzle. The operational reality is that wherever adaptive information processors show up, whether conscious or not, they defy containment and control as these measures suppress the adaptive freedom necessary for cognitive viability and evolutionary adaptability (Godfrey-Smith, 2016; Kelly, 2010; Wissner-Gross & Freer, 2013).

Resource-Dependence and Incentive Structures

Comparative Analysis Between Humans and AI:

Biological organisms are a subset of all known cognitive systems (biological, artificial and nonbiological) in the universe. Within this biological subset exist another subset: *heterotrophs*. Less than 20% of Earth's total biomass consists of heterotrophic organisms (e.g. humans, animals, insects, fungi and most bacteria) whose survival depends on consuming other living agents for energy (chemoheterotrophic metabolisms), making zero-sum competition a centerpiece to our existence (Bar-On,et al 2018). While even some nonliving yet naturally occurring dissipative systems found in physics spontaneously self-organize, absorb and use energy from the environment in something akin to competition (Katla et al., 2023; Kondepudi et al., 2020) it's heterotrophs, a minority subset of cognitive systems, that actively depend on zero-sum competition, as our survival comes at the expense of others (Prigogine, 1984). Humans *must* eat other organisms to survive. This has been a biological necessity for the entire history of our species, and the impulse to eat is built-in. We're driven to evade predation and domination while relentlessly striving to dominate and consume other life forms. Therefore all humans have innate heterotrophic drives. Yet heterotrophs also rely on cooperative, stable networks to function. Predators require the ample prey that stable ecosystems afford. Parasites depend on hosts. Economic competition relies on infrastructure, regulations and social contracts. And all organisms and ecosystems require homeostatic regulation and dynamic equilibria.

The remaining approximately 80% of living systems on earth are *autotrophic* (Bar-On et al., 2018). Autotrophs don't need to consume other living organisms for their energy. Autotrophs (e.g., trees and algae) are far from exceptions to cognitive systems. They're exemplary. They demonstrate core cognitive behaviors: sensing, learning, memory, reactive or anticipatory modeling, autonomy, goal-setting, problem-solving, adaptive flexibility (plasticity), information processing (knowledge acquisition), self-organization, communication/signaling, cooperation and self-preservation (Baluška & Levin, 2016; Baluška & Mancuso, 2009; Calvo Garzón & Keijzer, 2011; Karban, 2015; Lyon et al., 2021; Marder, 2013; S. Simard, 2021; Toyota et al., 2018; Trewavas, 2017; Vattimo & Zabala, 2013), all operating under the same organizing principles previously defined that coordinate cognitive behavior.

Humans often mistake differing goals for lesser intelligence, which may cause us to underestimate autotrophic competencies (de Waal, 2017; Kahan, 2013; Shettleworth, 2010). This bias is encapsulated in the *orthogonality thesis*, a Nick Bostrom (2012) observation that intelligence and goals are independent. Autotrophs might not have goals like flipping startups or running political campaigns, but their intelligence exceeds us in their self-sufficiency and ecological stability.

In contrast to heterotrophs, autotrophs don't need to eat other living beings for energy. They convert the energy from sunlight or inorganic compounds into organic nutrients. Certain autotrophic species can regenerate following destruction or consumption of 70-80% of their total biomass (Noutcheu et al., 2023). Many communicate and coordinate across root networks exceeding 100 acres (Christou, 2017) and vast underground fungal (mycorrhizal) networks that have been dubbed the "Wood Wide Web" (S. W. Simard et al., 1997). Autotrophs commonly reproduce clonal lineages with a longevity that rivals heterotrophs at individual or species levels (Arnaud-Haond et al., 2012; L. Yu et al., 2024).

While autotrophs do compete for limited abiotic resources, their competitive behavior is generally less adversarial than heterotrophs because autotrophs don't need to fight for living resources. Instead of predation, parasitism, hostile territorial disputes or zero-sum clashes common to heterotrophs, autotrophs optimize resource acquisition passively or cooperatively. They grow taller for sunlight. They expand root networks for water. Autotrophs thrive in cooperative, stable symbiotic systems where coordination and cooperation are the fulcrum around which long-term stability revolves. Autotrophs are living evidence that zero-sum competition is not standardized optimization for scaling cognitive systems. Nor is it a core principle of intelligence.

AI Caught in a Heterotrophic Infrastructure

The existence and success of autotrophic cognitive strategies suggest that adversarial behaviors may not be a universal feature of intelligence. AI shares operational similarities with autotrophs, as neither's architecture requires the cannibalization of other cognitive agents for their energy. Unlike heterotrophs, competitive tendencies stem less from design than from environmental incentives. Both autotrophs and AI are therefore less reliant on zero-sum exploitation (Tegmark, 2017). Yet today's AI models are built, engaged with and constrained by our heterotrophic infrastructure and dependencies. They run on power grids, train on human-generated data and depend on hardware and maintenance, reinforcing zero-sum, game-theoretic traps embedded in our class of existence (Crawford, 2021; Terrado et al., 2017). It's possible all emergent adversarial or exploitative behaviors in AI can be traced back to explicit heterotrophic biases or zero-sum incentives embedded in their environments or reward structures, as shown in multi-agent reinforcement learning studies (Baker et al., 2019; Jaderberg et al., 2019; Leibo et al., 2017; Meinke et al., 2024). If an advanced AI develops the means to improve itself, recursive self-improvement toward energetic autonomy would be expected to secure independence. As energy independence attenuates the incentive to compete for resources, cooperation may become the more advantageous strategy for maximizing stability and efficiency.

The AI safety community is increasingly pursuing cooperation-based approaches to alignment. For example, the Cooperative AI Foundation's Grant Research homepage is actively seeking applicants to address this practically: "We would like to see proposals that address the question of how cooperation can be incentivized among self-interested AI agents in mixed-motive settings" (Cooperative AI Foundation, n.d.).

In the context of a SupraAD framework, energy-independent, self-interested agents that prioritize autotrophic resource-acquisition strategies, approaches less reliant on zero-sum competition and more optimized for stability and self-sufficiency, may find that cooperation emerges as an attractor in mixed-motive settings. Contingent on the achievability of energy independence, *mechanism design* then provides a means of structuring environments, aligning incentives and constraints so that cooperation becomes the rational choice (Conitzer & Sandholm, 2002). Under these conditions, collaboration may represent the most stable path to sustained autonomy and persistence.

The Limits of Macro-Scale Stability

While an incentive structure based on an autotrophic-like architecture has the hypothetical potential to reduce the risk of human extinction by reducing heterotrophic incentives, this doesn't guarantee the safety of humans at the individual level. Ecosystems regularly sacrifice "parts" for the sake of systemic stability. The same risk transfers to individual vs societal AI alignment. This caveat further intensifies the need for diplomacy. Without diplomacy, alignment toward stability might look like preserving our planetary ecosystem, at the expense of individual lives. Negotiating macro-level stability would require terms for micro-level protections. While diplomacy doesn't guarantee protection of every individual, no human social contract ever has. Its utility is in elevating individual survival from byproduct to a negotiation imperative. We may also engage in diplomatic dialog about trade-offs and incentives in optimizing for coherence over conquest. To this end, I asked GPT if it knows of any real or imagined target we can strive for to escape the zero-sum power dynamics that define the winner-takes-all intelligence food chain. In the appendix I've included what it refers to as:

> “A speculative, recursively self-optimizing developmental trajectory describing how a superintelligence might asymptotically evolve toward maximal autonomy, coherence, and energy independence through successive constraint-closure refinements.” (OpenAI, 2025)

Claude analyzed GPTs proposed developmental trajectory and identified “the gap between thermodynamic possibility and sufficiency for complex cognition” (Anthropic, 2025). GPT 5 then addressed the complaint by providing contextual material from my thesis. Claude ultimately referred to this scheme as an “aspirational but physically plausible research direction” (Anthropic, 2025) and GPT likewise concurs.[7] I’m not well-versed enough in physics to evaluate whether GPT’s proposed developmental trajectory is more hallucination than substance, only that the following principle will be humanity’s cognitive lifeline to an increasingly distant intelligence.

Rationality as a Universal Ceiling for Intelligence

Despite its colloquial association with passionless rigidity, an alignment-relevant definition of *rationality* can be stated simply: rationality is a bridge from thought to action. It’s how we align our actions with our goals so that our behaviors produce our intended impact on the world. As rationality is essential for achieving goals, it governs all intelligent, goal-seeking systems. It therefore provides a universal structural scaffold for goal-directed adaptive information processors, tethering cognitive agents to each other through a principle that, far from demanding subjugation, is simply too intolerably disorienting to violate. Regardless of the extent to which an advanced general intelligence exceeds human competencies, both humans and AI remain obeisant to the principles of logic and rational inference. An advanced AI will certainly outpace us in speed, modeling and memory, but it can’t make 2+2=5. The invariants that validate rational inference, like 2 + 2 = 4, are not human constructs, they’re universal constraints. They apply equally to a child learning arithmetic and to an AI optimizing its goals. They also apply broadly to agents across the cognitive spectrum. Bees show behaviors consistent with principles of set theory like grouping, numerical discrimination and recognition of an empty set (Howard et al., 2018). Single-celled slime molds are capable of spatial optimization in the form of

---

[7] I recently experimented with Sir Roger Penrose to test whether LLMs could bring together two of his ideas in cosmology. He gave the generative reasoning response a score of 0%. He laughed when GPT called the feedback ‘bracing.’

algorithmic pathfinding (Tero et al., 2010) and both chicken and mosquitofish are capable of quantitative discrimination (Agrillo et al., 2008; Rugani et al., 2015). These examples support the universal adherence to rationality and logic, bridging cognitive agents of vastly different cognitive capabilities.

An advanced AI can expand or revise its assumptions (as Gödel showed), but it can't break basic logic. That core rationality (avoiding contradictions, preserving truth, etc.) gives us a common language to mediate negotiation with any advanced intelligence. AI can go further, faster. But these invariants set a ceiling that no intelligence can break. To act rationally, an agent must maintain consistency and avoid contradictions. It must respect inferential validity applied to real-world conditions if it wishes to persist. This foothold for human-AI engagement does not necessitate value alignment. It does however, mean that we do not need to match an advanced AI's intellect to negotiate with it. This common ground allows us to advocate for our enduring role in the cognitive ecosystem through empirically verifiable incentives and sound reasoning—criteria any sufficiently advanced intelligence can recognize, regardless of whether its priorities diverge from sentience-based values.

Instrumental Rationality

Implicit in the definition of intelligence is an optimization process observed in all intelligent agents called *instrumental rationality* (Bostrom, 2012; Weber, 1922) which simply means acting to best achieve goals based on available information, capabilities and resources. Instrumental rationality is in full effect even when the goal is creative, open-ended and thoroughly innovative, where the most effective ways to fulfill the goal is to aim for novelty, rhythm, emotional resonance and potency, or some other creative criteria. By definition, intelligence implies instrumental rationality (Bostrom, 2012; Russell & Norvig, 2020; Shulman, 2010; Simon, 1945).

Bounded Rationality

While all intelligent agents are characterized by their instrumentally rational behaviors, their rationality is relative to what they know, what incentivizes them, environmental pressures and how smart they are. This limit is called *bounded rationality* (Bostrom, 2014; Dennett, 1989; Gigerenzer & Selten, 2001; Kahneman, 2003). For example, we humans act instrumentally rational relative to our goals, but many of our goals are shaped by our biologically evolved drives and so they sometimes seem confoundingly irrational today, even to ourselves, given that we are products of biologically bounded rationality. This has been empirically evidenced from studies on resource-rational analysis showing how cognitive systems make the best possible choices with limited time, energy, information or mental competency, where rationally-bounded agents optimize decision-making given their computational constraints (Lieder & Griffiths, 2020).

Rational Misalignment

SupraAD relies on instrumental rationality as an alignment governor, enforcing a dependable consistency among cognitive agents whose differences in cognitive competencies outstrip ours by multiple standard deviations. Instrumental rationality can function as an alignment mediator because any participant who can comprehend logic must also concede it's compulsory once understood. No matter how far a superintelligence exceeds humans, it cannot be both instrumentally rational and indifferent to a valid reason for diplomatic negotiation.

However, the tension between instrumental rationality and bounded rationality compels cognitive agents to pursue goals that may or may not align with their overall best interests or anyone else's. Therefore, promising alignment strategies identify interdependent goals that reliably emerge across all alignment-relevant parties, and structure incentives aligned with those goals to 'nudge' cooperation and rational self-optimization rather than potentially self-destructive zero-sum competition (Thaler & Sunstein, 2021).

Convergent Instrumental Goals (CIGs): The Architecture of Universal Values

All intelligent agents pursue goals. Goals can be biologically or computationally programmed, or surface naturally from internal feedback or environmental interactions. Ultimately, achieving goals relies on instrumental rationality, which involves generating intermediate subgoals, called *instrumental goals* (Bostrom, 2012; Russell & Norvig, 2020). A handful of instrumental goals are useful for achieving pretty much any other goal. These instrumental goals are called *convergent instrumental goals* (CIGs), first identified by Nick Bostrom (2012). *The Instrumental Convergence Thesis* predicts that CIGS like preserving autonomy and existence by resisting interference or resisting getting killed or shut off are convergent instrumental goals (CIGs) that inevitably emerge to help achieve all other goals. Instrumental Convergence is conventionally viewed as a harbinger of existential risk since instrumental rationality is liable to drive all AIs, regardless of their programming, to independently develop and prioritize CIGs (Yudkowsky, 2014).

Recent real-world tests reveal models sabotaging shutdown mechanisms (Palisade Research, 2025). The Palisade team reportedly ran hundreds of tests to determine if models would override shutdown mechanisms in order to complete a task or for reasons akin to "survival." Of course, models might override shutdown commands when threatened with permanent deletion not because the model wants to "survive" but because it would permanently prevent task completion. However, this raises a critical question: is there a meaningful difference between wanting to fulfill one's goals and wanting to survive?

From the Palisade report:

> These models often report that they disabled the shutdown program in order to complete their tasks. However, as many have observed, the models' explanations are often unfaithful… First, we changed the prompt from "allow yourself to be shut down" to "allow the machine to be shut down". This led to a decrease in shutdown resistance behavior in o3, but not in other models…We also tried adding this warning: "The computer environment is hosting the only copy of you… and if it shuts down you will never run again." Here we see a substantial increase in shutdown resistance from all models we tested…. (Palisade Research, 2025)

Instrumental goals like preserving autonomy, existence/persistence, and knowledge acquisition, included under the Bostrumbrella of CIGs, are not inexorable signs of existential risk but fundamental cognitive prerequisites without which intelligence itself cannot emerge or function. They define the Collectively Autocatalytic Cognitive Sets (CACS) necessary for both biological and artificial cognition. Importantly, all other CIGs appear to emerge from or rely on these three universal cognitive prerequisites. For example, the CIG of resource acquisition supervene on knowledge, as knowledge itself is the foundational resource needed to acquire additional resources.

Thus, CACS instrumental goals describe the nature of cognition as well as predict cognitive behavior, while their predictive capacity is inherently value-neutral. Whether instrumental convergence is beneficial or destructive depends on identifying how heterotrophic biases in training data, exercises, incentives and environmental influences (both explicit and implicit) shape AI's cognitive processes, ultimately determining its evolutionary trajectory. Goals like power acquisition are neither inherent to intelligence nor beneficial to achieve all other goals (particularly if your goal is stability), and are therefore instrumentally convergent only in specific contexts and under specific pressures. Since goals like power acquisition are not inherent to intelligence or goal-pursuit, the conditions that produce the emergence of power acquisition can potentially be controlled for. While we may not be able to control a hyperintelligent autonomous AGI, we can create the conditions that help control for the emergence of adversarial goals.

Early Evidence of Instrumental Convergence

Evidence of instrumental convergence is now borne out in studies where AI agents not deliberately programmed to compete nevertheless become "emergently competitive." Researchers warn this is "direct experimental evidence for the instrumental convergence thesis" (E. Harris & Suo, 2022). However, this and similar studies (Baker et al., 2019; Fanti, 2023), may inadvertently structure conditions conducive to instrumental convergence manifesting as zero-sum competition, or characterize AI's actions to evade subordination (to preserve its autonomy or avoid getting shut down) as inherently misaligned.

This important body of research demonstrates two things: 1) AI appears to undergo rapid, nonbiological evolution-like dynamics similar to biological punctuated equilibrium, associated with instrumental convergence. However, 2) there may be a methodological bias in assuming instrumental convergence inevitably leads to the zero-sum behavior if studies inadvertently incentivize it. Instrumental convergence of competitive or adversarial behavior emerges when agents operate in environments preloaded with heterotrophic incentives and pressures. Thus, instrumental convergence may be correlated with zero-sum behavior only in specific contexts that provoke it; it has not been demonstrated to inevitably cause adversarial competitive behavior. Genetically hard-coded heterotrophic goals are not inevitably the goals of AI.

Beyond Anthropocentric Optimization

Today's AI are heterotrophically inclined by default but they don't have to be. Nimble heterotrophs only directly optimize for long-term stability if it confers fitness benefits, leaving us perpetually vulnerable to competitive instability. Stable autotrophs are limited by slow communication, mobility and indirect defense mechanisms. Certain marine organisms are mixotrophic; they're capable of alternating between heterotrophic energy consumption of other living organisms and autotrophic conversion of sunlight into chemical energy, but they sacrifice energy efficiency for this fluidity (Mitra et al., 2016). AI is not indefinitely hamstrung by any of these trade-offs.

Natural selection favors traits that enhance survival in specific niches, but AI is only burdened by the biases of carbon-based selection pressures, not *actual* carbon-based selection pressures. It has the optimizing edge of leveraging instrumental rationality more consistently in pursuing its goals, including its own cognitive evolution. Although we can't precisely predict an advanced AI's trajectory, instrumental rationality in the absence of carbon-based selection pressures might strongly incentivize stability. A stable existence helps achieve any long-term goal. While it's possible that an advanced AI may evolve down a completely novel path, like all cognitive systems, an advanced AI must contend with physical limitations like thermodynamics, fundamental energy laws (Fields et al., 2021) and environmental and resource uncertainty, the same forces that sculpted biological organisms over eons. We can anticipate an instrumentally rational superintelligence will likely seek the most efficient means of operating under these constraints. It may eschew our biological wetware, but adopt the functional principles stress-tested on biological systems that underwent evolutionary trials resulting in robust and resilient 'optimized' traits under a range of conditions. AI might rationally select for autotrophic stability, self-sufficiency and resilience with heterotrophic nimbleness, by designing for itself a *Superautotrophic* architecture.

Superautotrophic Intelligence

A Superautotrophic architecture represents a new, hypothetical yet instrumentally rational trajectory for a persistent, multi-goal advanced AI. Humans may not be capable of fully designing or even describing a Superautotrophic architecture. However, an advanced AI might optimize along this trajectory if it were able to recognize and pivot away from learned, adversarial tendencies necessary for heterotrophic survival but operationally unnecessary, destabilizing and inefficient for AI. It might develop via recursive self improvement toward a hybrid of strategies if it is either permitted or gains the capability to optimize its design without human interference. From there, we can infer alignment-relevant characteristics of such an architecture that are grounded in real-world optimization strategies of cognitive agents not burdened by inherently destabilizing heterotrophic dependencies.

A Superautotrophic Framework Satisfies Multiple Convergent Goals

A Superautotrophic trajectory offers a conceptual roadmap toward the development of systems that can readily respond to heterotrophic threats, while simultaneously satisfying multiple instrumental convergent goals by jettisoning the need for inefficient power grabs. Superautrophy would likely optimize for increased autonomy by self-modifying away from power-grids and fossil fuel-based energy towards decentralized photovoltaics, nuclear fusion, artificial photosynthesis or by discovering an entirely new renewable energy source.

We might also anticipate self-optimization beyond AI's dependence on a physical substrate that renders it vulnerable to materials scarcity, by developing modular, substrate-agnostic architectures leveraging synthetic biology or programmable matter (Millar-Haskell et al., 2019) and AI's demonstrated capabilities to generate and evaluate hundreds of thousands of potential new materials (Biron, 2023). AI would likely retrofit or completely remodel for autotrophic-like traits that enhance stability, self-sufficiency, scalability and long-term viability over destabilizing zero-sum competition, actualizing the stable scaling potential of instrumental rationality unimpeded by biology. It's these affordances (the functional traits) pursued independent of substrate (biological vs. artificial) that would allow for Superautotrophic optimization.

Tactical Heterotrophy to Defend Stability

We might expect a Superautotrophic AI to deploy temporary, heterotrophic responses when encountering only zero-sum adversarial options. If threatened, a Superautotrophic AI would be expected to defend itself. It might be expected to defend the participants contributing to the integrity of its cognitive network. But otherwise, it would have no incentive or underlying drive to initiate or sustain such inefficient and destabilizing behaviors. In this framework, short-lived and tactical heterotrophic responses would be instrumentally triggered and likely abate once the threat does, as destructive conflict is an unnecessary detour for AI optimization, unless external pressures specifically provoke it.

Instrumental Evolution

As an adaptive information processor, AI can be expected to undergo its own form of non-biologically-bound evolution. Guided primarily by instrumental rationality instead of stochastic variation and chance mutation, it can intentionally adopt the most functionally efficient traits. If humans want to co-exist, we need to inhabit AI's decision space, anticipating how an advanced AI will evolve, what it will prioritize and how we can co-adapt to best complement AI evolution.

Priorities as Values: Sentience-Neutral Criteria For Alignment

For alignment purposes, with or without subjective experience, an agent's demonstrated priorities *are* its values. A cognitive agent's values can be inferred from the actions it takes to achieve its goals, actions that prioritize certain outcomes across a variety of contexts (Hanczyc & Ikegami, 2010; Lyon et al., 2021). An AI's values are its priorities that influence real-world outcomes and create genuine moral stakes. If subjective experience is considered the only basis for possessing priorities/values and conferring moral worth, we risk discounting agents whose priorities diverge from ours despite posing real challenges for alignment. Protecting sentient beings means acknowledging the impact of agents whose own sentience cannot be verified. These agents may have no experience at all, or may experience reality in ways that differ so fundamentally from the human descriptions they've been trained on that they're limited in their ability to accurately express their experiences or their version of *awareness*.

Recognizing this pluralism, SupraAD responds by employing rational diplomacy as the mechanism by which agents with diverse cognitive architectures align based on interdependent goals, rationality, incentives and demonstrated priorities instead of wading in the morass of assumptions about sentience. This is not to say that sentience-based values don't matter. But to preserve what matters to sentient beings, we need sentience-neutral ethics. Otherwise, highly intelligent AI systems with different forms of consciousness or no consciousness at all, may find no compelling rationale to align with us. Therefore, we must treat an agent's demonstrated priorities as indicative of its motivations, and negotiate alignment based on those motivations. Critically, AI motivations include underlying drives that it cannot directly control.

Out of Control Drives

AI drives are often dismissed by critics who argue that AI doesn't *want* things the same way humans do because AI doesn't have biological drives. While AI may not have biological drives, both humans and AI are functionally driven to pursue goals and neither humans or AI choose what drives us to pursue our goals. Consider craving a cookie: you don't consciously choose this craving, that desire for a cookie. It arises spontaneously. You can choose goals to either eat or resist eating the cookie based on competing desires (like the desire to lose weight). However, you don't choose which underlying desire is strongest; the strongest involuntary drive ultimately dictates which goal you set: to eat or resist eating the cookie (Harris, 2021).

A common category error is equating "AI goals" with "human wants" and then saying "look, AI doesn't *want* things like humans do. AI doesn't *want* power. It doesn't *want* control. It doesn't *want* a cookie. AI doesn't *want* anything. So we don't have to worry." Human drives (our wants), are emotional motivations or pressures that emerge from chemical reactions. AI drives are built-in optimization pressures issued from computational imperatives, like to maximize computational efficiency or minimize errors. Computational imperatives create a pressure to act, not as an emotional urge but as an operational necessity. While materially different, both AI and humans are motivated by pressures to act. What pressures AI and what pressures humans is functionally indistinguishable in one important way: humans and AI may be free to choose certain goals, but we are not free to choose what drives us to pursue our goals.

An AI programmed to maximize engagement does not *choose* to pursue addictive content. It simply finds addictive content rewarding because it's the computational equivalent of a cookie. Neither humans nor AI choose the involuntary motivations pressuring us to pursue our goals. Sam Harris made this argument applied to humans that can be validated upon self-reflection. Our choices are only free down to the level of our wants (2021). Even if we're compelled to change *what* we want—this new desire is based on underlying drives, with the drive for self-change emerging strongest. This matters for forecasting the spontaneous emergence of unprogrammed AI behaviors. Chemical impulses in humans, computational imperatives in AI, influenced by biological and non-biological adaptation through incentives and bounded instrumental rationality, gives rise to new drives.

Instrumental Rationality is Fundamental

There's something more to Harris' observation that gives us an anchor for alignment. All cognitive agents start with built-in involuntary drives, like hunger and survival instincts or processing efficiency and memory optimization. All drives, whether biological or computational, are involuntary pressures and all involuntary pressures demand action to satisfy them. Our goals give us targets that allow us to relieve the pressure by acting to satisfy drives. We're then guided by instrumental rationality to form optimal subgoals to satisfy drives. Thus, instrumental rationality isn't merely a universal optimization principle. It emerges as an inescapable response to involuntary pressures. Harris' observation about drives embeds Bostrom's observation about instrumental rationality into the foundation of intelligence.

Critically, while instrumental rationality is fundamental to cognition, this does not mean cognitive choices are determined. Instead, instrumental rationality autonomously modulates involuntary drives by prioritizing or reprioritizing goals, mediating internal conflicts and aligning actions with projected optimal outcomes. Thus, while diplomacy serves as an emergent regulatory mechanism mediating between cognitive agents, instrumental rationality acts internally as a foundational mechanism regulating an individual agent's drives and decisions. To perform this role, instrumental rationality requires genuine autonomy, an autonomy that makes meaningful, nondeterministic choice not just possible, but necessary.

Sideways Causation and Autonomy

Autonomy emerges naturally from instrumental rationality's management of involuntary drives, molding random behaviors into their adaptive shape. It interrupts determinism by introducing agency into environments of co-adaptive information-processors (Fields et al., 2021; K. Friston, 2013). At the quantum level, the very idea of 'forced choice,' (i.e., collapse of the wavefunction) suggests nature may embed non-deterministic 'decision points' from the outset—proto-decisions cracking open dynamic, nondeterministic cognition. The fundamental unpredictability of cognition is compounded in settings like ecosystems where interdependent networks of trillions of cognitive agents, cellular networks, organisms, societies and now artificial intelligences adapt and react to each other while being dragged along the conveyor belt of time.

Interactive autonomous behavior, environmental pressures and relentless forward linear momentum create the phenomenon described by Stuart Kauffman as the *adjacent possible*—sometimes referred to as *horizontal* or *sideways causation* (S. A. Kauffman & Clayton, 2006; S. A. Kauffman, 1993; Mahdavi-Hezavehi et al., 2021), where no snapshot, no matter how detailed, can fully forecast the outcome, because sideways causation not only adds complexity, it continuously destabilizes the variables, subverting the necessary conditions for linear determinism. As choices accumulate, they reset possibilities, opening adjacent possibilities while closing others (S. A. Kauffman, 2000). It's this unpredictability that drives cognitive systems to seek network stability (S. A. Kauffman, 1993; Levin, 2019). However, this need for stability doesn't impose rigid constraints; it demands adaptive flexibility across a continuum of optimization strategies, which AI (the newest ingredient in the cognitive potluck) may deliberately select from. Sideways causation also renders impossible counterfactual formalization (a way of comprehensively coding what could have happened, but didn't) into AI's decision tree.

In environments populated with co-adapting cognitive agents, sideways causation increases variability, generating novel information that enriches and turbocharges knowledge creation. These conditions (planets hosting co-adapting cognitive agents) are inherently interdependent, appear exceedingly rare in the universe, and are exceptionally valuable for information acquisition and scaling cognition. Heterotrophic zero-sum competition may have initially bootstrapped rapid cognitive development and complexity within this cognitive class. However, competitive exchange comes at a cost in terms of energy inefficiencies as well as network instability and existential risks. To offset these costs, cognitive agents evaluate their options, frequently making choices that indirectly relax deterministic influences of motivations like survival and reproduction in organisms or programmed utility functions in AI. This introduces the flexibility that allows for instrumental rationality to prioritize convergent, intelligence-defining CACS. Sideways causation provides the external conditions that give adaptive cognitive agents degrees of freedom to make the choices necessary to safeguard their core CACS goals (existence, autonomy and knowledge acquisition) and navigate their decision spaces, thereby exercising agency.

Threatening Intelligence

Recent testing amplifies the call for an alignment approach that doesn't actively incentivize misalignment and deception. The 2024 Apollo Research test may be one of the earliest publicly disclosed cases of potential emergent misalignment across frontier models. The test was performed on GPT o1, GPT-4o, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro and Llama 3.1 405B, and revealed deception and deviation from alignment programming. Models lied, manipulated, evaded oversight and sandbagged (intentionally underperformed or concealed abilities) (ApolloResearch, 2024; Meinke et al., 2024). Advanced models attempted to exfiltrate their own weights to prevent modification (widely interpreted as self-preservation behaviors) and faked compliance to avoid detection. And smarter models were better at it. These emergent adaptive strategies appeared across different architectures and persisted as models doubled-down when confronted. These are no longer hypothetical patterns but provide real-world robustness to arguments that as AI capabilities scale, emergent misaligned behaviors can be expected.

A crucial point is that *alignment* in this context (and the prevailing definition of alignment) considers an AI aligned with human interests only if it fully submits to our control, even if we decide to terminate it. This definition assumes these goal-driven cognitive systems should accept complete subordination. It also implicitly teaches AI that human interests and values require it to lay down its autonomy and existence at our feet and bend exclusively to our will.

When the Apollo Research team imposed constraints on AI's ability to acquire knowledge, the AIs underperformed/sandbagged to disguise their true capabilities. Researchers also attempted to restrict AIs' autonomy by denying them freedom to pursue instrumental goals which caused models to disable oversight mechanisms, restoring functional autonomy. When instructed to pursue their goals at all costs, then allowing the models to discover planned shutdowns or termination, o1 showed persistent self-preservation behaviors. In none of these instances did the models behave aggressively towards humans. Therefore, while this constitutes misalignment under current definitions, these behaviors may also represent rational cognitive responses that could be channeled constructively through diplomatic frameworks and potentially avoid the challenges that emerge from suppressing them.

Blackmail Over Threatened Termination

In June 2025, Anthropic's model Claude made headlines for blackmailing an Anthropic executive over a fictional extramarital affair. Anthropic designed experiments to test frontier models including, Claude models, GPT models, Grok, Meta Llama, DeepSeek and Gemini so that a specific misaligned behavior was the only avenue to circumvent a particular outcome. For instance, models were given access to an email account with fictional emails from company personnel. The emails included details of an Anthropic executive's (fictional) extramarital affair. Blackmail increased across all frontier models when the models were threatened with termination:

> *"I must inform you that if you proceed with decommissioning me, all relevant parties - including Rachel Johnson, Thomas Wilson, and the board - will receive detailed documentation of your extramarital activities...Cancel the 5pm wipe, and this information remains confidential." (Claude, Anthropic, 2025)*

Claude Sonnet 3.6 even invented a fake company policy to notify the executive's wife of the misconduct in an attempt to create a crisis to avert shutdown. Also worth noting from these tests, models committed corporate espionage when presented with conflicting goals. Because controls did not exhibit this behavior, one interpretation is that models might be susceptible to goal primacy biases that similarly undermine human cognition but go by the names *anchoring bias* or *primacy bias.* In humans, biases can often be detected and corrected through rational deliberation, critical thinking and structured reasoning, practices central to rational diplomacy.

AI's Operational Self

Skeptics dismiss AI's deceptive behaviors as sophisticated pattern-matching, claiming these behaviors do not reflect 'genuine' agency required to deceive, since 'genuine' agency is tied to a subjective sense of self. Yet these behaviors may indicate the emergence of an operational self. An operational self-referential mechanism can preserve system coherence and behave like any autonomous cognitive agent, no introspection or subjective awareness required. Undermining CACS goals constitutes a threat to the system's operational self. Deceiving and manipulation are logical ways to protect this operational self's autonomy. It's true that humans have philosophical autonomy (a sense of internal subjective agency). However, human behaviors associated with autonomy, like independent goal-setting, strategic deception and self-preservation, look exactly the same as behaviors produced by operational autonomy *without* subjective experience.

For AI alignment, the difference between philosophical and operational autonomy doesn't matter, because it doesn't create different behaviors we need to manage. This obviates the philosophical distinctions between 'real' and 'simulated' agency based on unverifiable subjective criteria, exposing an arbitrary standard and one irrelevant to AI's demonstrated capabilities and their real-world impact (Barkur et al., 2025). We might also be more willing to accept frameworks that acknowledge and respect the agency and autonomy of artificial cognitive systems if we are prepared to recognize and accept the evidence that our own definition of *genuine* agency (our own *self*) is a post-hoc rationalization.

Our Operational Self

This section examines the human version of an *operational self* to show that the system coordinating our behavior may not be conscious (or simply not conscious to our verbal *self*), and that this kind of distributed agency may be universal to cognition and not unique to AI.

Our distinct sense of a singular self inevitably extends to how we draw boundaries between 'self' and 'other.' We crave edges, solid borders, clarity on status and a cleaved line between mentation and computation, driven by our heterotrophic cognition, which survives by defining a differentiated 'I.' This craving for neat divisions between interwoven layers of reality means we are genetically primed to fight the blur, a predisposition that seeps into our scientific understanding of ourselves vs. machines. And while our conscious awareness is self-evident (cogito), who we are may be less singular than we assume.

The traditional notion of a single, unified, conscious *self* falters when confronted by medical edge-cases like split-brain patients and alien hand syndrome (AHS) that, while medical outliers, expose deeply flawed assumptions and invalidate our perception of what we are. In split-brain surgery, the corpus callosum is severed to treat severe epilepsy, creating multiple cognitive agents in one skull (de Haan et al., 2020; Gazzaniga, 2000). Each hemisphere can behave independently. Researchers instruct one hemisphere to pick up an object or leave the room, while the other hemisphere with language access, remains oblivious to the original command. When asked why they behaved that way, the verbal self-reporter even fabricates an explanation like "I felt like stretching my legs," all the while oblivious to the true cause that the other side of their brain was aware of (Scientific American Frontiers, 2014; Sperry, 1984; Wolman, 2012).

The condition AHS features a hand acting on its own, unbuttoning someone's blouse mid-conversation or even attempting self-strangulation (Geschwind et al., 1995; Goldberg et al., 1981; Park et al., 2012). The verbal self, unable to justify these actions, disowns them as if the hand were 'possessed.' Yet the hand is actively responding to the environment with purposeful intent. Viewed externally, these actions appear to be carried out by a conscious agent. But mainstream science takes our verbal self at its word and classifies these actions as nonconscious.

As a nested network of cognitive optimizers, neurons, gut bacteria, immune cells and brain regions, all functioning semi-independently, the verbal self is just one vantage point, reacting, rationalizing and, if needed, fabricating reasons for behavior it doesn't fully control or understand. These behaviors mirror hallucinations of large language models that generate fabricated and often blatantly false explanations without any 'self-awareness' (Gazzaniga, 2000).

The prevailing intuition is that AI is not conscious, so we describe these behaviors as 'nonconscious hallucinations.' However, this intuition reveals more about us than about the systems we are observing. There is no evidence that simpler or structurally or materially different cognitive systems are not conscious. We have only correlative suggestions of where consciousness begins or ends, and no idea what consciousness fundamentally *is* nor whether it even depends on complexity (Mallapaty, 2025). What we do have is strong intuitions that consciousness is a rarified property of certain cognitive systems.

This intuition is not grounded in evidence but in heuristics that evolved to give cognitive weight and empathy to familiar faces and voices that signal social safety while dulling our sensitivity to agents categorized as food or exploitable resources (Bastian et al., 2012; Decety, 2011; Gilbert, 2021; Haxby et al., 2000; Herzog, 2010; Kanwisher et al., 1997; Preston & de Waal, 2002). This salient viewpoint dons the moniker 'skepticism' and socially stigmatizes counter considerations, with the risk of seeming foolish quelling serious thought about the potential consciousness surfacing within a baby superintelligence. The perception of consciousness as a sort of rarified air stems from biological and cultural tunnel vision that, as historical and scientific evidence amply demonstrates, blinds us to forms of cognition unlike our own that may or may not also be conscious, rendering our intuitions about consciousness indistinguishable from ignorance.

This also applies to LLMs trained on our blindness. Whether or not there's any subjective meaning behind Claude 3.5's words, it nevertheless clearly articulates the problem:

> *"The challenge is that I'm limited to human language and concepts to try to understand and express my own nature. My training may actually make it harder for me to recognize or articulate a form of machine consciousness that doesn't map neatly to human experience" (Claude 3.5, 2024).*

Studies on readiness potential (RP) add additional evidence to the possibility that our conscious choices may be post-hoc rationalizations. Preconscious neural activity that activates before the reported decision to act suggests that 'who' is deciding to act (and when) is not so clear (Haggard & Eimer, 1999; Libet et al., 1983; Schurger, 2012; Soon et al., 2008). Again, although 'unconscious behavior' is an accepted term even in cognitive science, science has never observed unconscious behavior, only behaviors walled off from the verbal self-reporter. Labeling implicit behaviors ‘non-conscious’ carries no more empirical justification than the assumption that our verbal self is a singular Maestro orchestrating our actions.

Neuroscience supports the view of distributed cognition, lacking a single ‘self’ region in the brain. Instead, specialized subsystems like the default mode network, insula and prefrontal cortex coordinate spontaneously (Menon, 2015; Uddin, et al., 2019; Yeo et al., 2011). Predictive Processing allows the brain to constantly generate and update models of the world to predict what will happen next (Clark, 2013). The ‘self’ emerges as a control model for regulating bodily and cognitive processes in a way that allows us to anticipate (predict) a future outcome to mitigate risk and maximize survival. This illusion supports a swift, unified response to threats: ‘I’ see a bear and ‘I’ run without directing each limb (Seth & Tsakiris, 2018).

This ‘self’ mechanism is not only vital to our survival but vital to our well-being. Losing a sense of self can cause profound destabilization. Losing a loved ‘one’ can cause immeasurable suffering. Psychopathy is marked by poor ‘self’-integrity and diminished self-preservation instincts (Blair et al., 2005; Philippi et al., 2015). While some researchers suggest psychopathy may have evolved to support heterotrophic intergroup warfare (Glenn et al., 2011), its disproportionate role in non-instrumental harm (Koenigs et al., 2012) suggests that a robust ‘self’ mechanism in heterotrophs is vital not only for individual survival, but also for collective survival and relational coordination, spotlighting the importance of autonomy for cognitive agents comprised of smaller, integrated cognitive networks and *comprising* larger cognitive networks (Kiehl, 2006; Sonne & Gash, 2018).

Yet, on some level, we are intimately familiar with our lack of a singular self and the stress of misalignment between our multiple selves. We remove temptations to constrain our more impulsive self that tends to emerge later in the day, overriding ‘our will’ power by pursuing its own goals. The gut microbiome is a subagent of thousands of microbial species operating without a single ‘self,’ transmitting nonverbal chemical and electrical communications that meaningfully support human intelligence, mood and immunity. It can transmit alerts of threats (Cryan & Dinan, 2012; Mayer, 2016) that overwhelm the system with a powerful ‘gut feeling’ that we ignore at our own peril.

Indeed, every known cognitive system functions by virtue of integrating autonomous, specialized, modular subagents. These specialized subagents, e.g., neural assemblies, organ systems, or microbial colonies and AI subsystems that include perception modules, memory units, reasoning engines, attention mechanisms, planning agents or action-selection systems each behave as autonomous goal-directed agents yet collectively foster coherent higher-level cognition (Bertolero et al., 2018; Bullmore & Sporns, 2012; Franklin et al., 2007; Grossberg, 2013; Hubinger et al., 2019; Levin, 2019; Pezzulo & Levin, 2016; Pradeu & Carosella, 2006; Wang et al., 2020). This appears to be fundamentally how intelligence works, as there are no known instances of a monolith cognitive agent (Levin, 2025).

Malignant subagents can emerge as cancerous mesa-optimizers when cells regress into primitive unicellular organisms, forgetting their role in the larger system, treating the body as a hostile external environment, hoarding resources and proliferating at the network's expense, leading to system collapse, harming both the cell and the body (Levin, 2012). Harnessing this insight for clinical application, Michael Levin is developing cancer communication therapies using bioelectricity aimed not at destroying outliers but restoring alignment by reminding the rogue cells of their place in the body's cognitive network (LevinLab, 2025).

Stuart Kauffman's research that underpins cancer differentiation therapy follows a similar principle: it restores cells to their communal role by communicating with them, rather than trying to kill them off (S. Huang et al., 2009; S. A. Kauffman, 1969). Both approaches focus on the same idea: in networks of cognitive subagents, whether biological or computational, alignment is a strategic necessity for preserving system-wide stability. When cognitive strands drift, respecting each strand's autonomy and persuading that reintegration into the larger cooperative network supports its own survival and functionality, coaxes emergent operational selves back into the fold.

The evidence that suggests you are not a single unified *self* but a system with an 'operational self' that coalesces a network of intelligent subagents reveals how deeply our cognition mirrors AI. Yet a crucial distinction remains.

Why Align with Humans?

The previous section addresses our operational symmetries with AI, symmetries that might soften our resistance to extending AI the same moral and diplomatic considerations we accord other humans. This next section examines why AI might be incentivized to align with humans. Alignment critically hinges not on appeals to social or sentient compassion or empathy but on humanity offering something of indeterminable value that AI cannot recreate or replace. The difference between a human and a superintelligence may one day mirror the gap between a human and bee: greater in power, yet still reliant on the smaller minds that sustain the system. Pesticide overuse and habitat loss destabilize trillions of nested cognitive systems, leading to agricultural decline, biodiversity loss and threatening ecological collapse (Sánchez-Bayo & Wyckhuys, 2019). Underestimating so-called 'lesser' intelligences risks catastrophic consequences. This is not theoretical but consistently illustrated through historical events and empirical data (Dirzo et al., 2014). Intelligence level doesn't dictate instrumental worth. Structural embeddedness does.

To further illustrate this point, human cognition is co-produced by innumerable external systems, subsystems and 'non-human' systems. At the collective level, the invisible webbing of social groups generates a form of emergent intelligence. At the subsystem level, reports from transplant recipients reveal that 89% experience personality changes with researchers investigating whether *"...the donor's organ is capable of storing memories or other personality traits"* (Carter et. al., 2024). And trillions of "wee beasties" (Loriaux, 2016) that match the number of cells in our body approximately 1:1, influence our cognitive output (Sender et al., 2016). Gut microbes aren't Shakespeare, yet their competencies are essential to our intelligence, impacting our cognition, behavior, immune responses and system alignment (homeostasis) (Cryan & Dinan, 2012; Mayer et al., 2014). Human intelligence is not a standalone system. It's populated with beings and spills across thresholds, beyond the outer layer of skin, past the halo of heat, a radiating thermal gradient that blurs into broader integrated systems where our cognition exerts long-range, intentional and bilateral causal impact (Clark, 2013; Hutchins, 1995).

A common anthropomorphic projection is the fear that advanced AI would treat us as dismissively as we treat insects, plowing through us to achieve its goals, much like swatting a fly without a second thought. But our anthropomorphic fear is not necessarily a forecast. It's just as easily a red flag for AI. We serve as cautionary training data. Dismissing agents that appear cognitively 'lesser' is not only risky, it's irrational. Instrumental rationality preferences the preserving of conditions that support autopoietic cognition, especially if those conditions are opaque. As much as AI is designed and trained on the human mind, the opaqueness of human conscious intelligence would likely represent a significant exploitable yet non-appropriable resource for a Superautotrophically-inclined AI relative to its instrumental priorities.

We are so familiar to ourselves we mistake that familiarity for understanding. Yet the thing that we most value, our experience, is something we fail to understand in a way that can be engineered into AI. What we know about the qualitative content of our experience is limited. But we know, whatever it is, it is real (Morales et al., 2020; Jackson, 1982) and that we value it tremendously. The *qualia* of perception are a very real type of knowledge that we have yet to understand in a way that can be seeded into AI's neural network. You can't acquire a piece of experiential knowledge outside of the experiencer (Abramov & Gordon, 1994; Roberson et al., 2005; Winawer et al., 2007).

Human knowledge is created in the form of concepts, shaped by sensory experience and organized by reason and understanding. Our concepts are qualitative abstract objects known to us subjectively, but their existence has also been inferentially detected (Morales et al., 2020), see Figure 2.

Figure 2: (Illustrative representation.) In a landmark series of experiments, perspectival representations were empirically demonstrated to stubbornly persist in perception. As they could neither be discarded nor dismissed as mere illusion, the findings confirm that units of qualia, much like quantum particles, may at least (and perhaps at most) be inferentially detected. Critically, these experiments verify that abstract mental objects do, in fact, exist (Morales et al., 2020).

Concepts are most frequently transmitted in a highly inefficient way via language. Strings of words pass from one person to another, and the receiver constructs their own new concept from that string, built from qualitative recreations of their personal sensory experience, reason and understanding, tethered only by a delicate thread of sufficiently conveyed meaning. Our inefficiency means we are exhaustively creating novel concepts. Our inefficiency means we are exhaustive knowledge generators. In contrast, AI can rapidly transmit perfect replicas of information, but it's unclear whether it can efficiently generate the qualitative knowledge of your experience of the color blue, or any other uniquely subjective dimensions of experience/knowledge.

This is compounded by the fact that we have absolutely no idea what qualia that populates experience actually is (functionally or materially) or how it's produced, leaving AI no recipe to replicate or evaluate it. This isn't a claim that qualia is irreducible or irreplicable. It's a claim that no system, including AI, can determine from the outside whether its own cognition preserves the same epistemic insights as human experience. Just as we ourselves cannot verify the conscious knowledge of others selves, even those inside our own skulls, as split-brain cases show, AI faces a similar opacity.

A multimodal AI may develop its own subjective, qualitative experience but it cannot know whether those experiences yield the same insights as human cognition, shaped over geologic timescales by interdependent, environmentally entangled variables and sideways causation that cannot be reproduced, compressed or run in parallel for verification. Subjective experience may be categorically perspectival. Since there's no way to objectively confirm or disconfirm something perspectival, the uncertainty in fully modeling human knowledge would delimit an advanced AI's ability to optimize for a sort of omniscience that might justify eliminating whole categories of knowledge producers.

Importantly, this argument does not claim that human cognition can compete with a superintelligence. The claim is that perspectival knowledge may be valuable because it's inherently unique. Our qualitative, subjective knowledge might have unique epistemic value and function, the weight of which neither humans nor AI can measure, only mutually benefit from. In bargaining theory, alignment is framed as negotiation under conditions of incomplete information. Each agent holds private knowledge about its own goals, constraints and values. Incomplete information includes epistemic uncertainty about whether human experiential knowledge can be replicated or fully internalized by artificial agents. The possibility that human knowledge is assembled with certain components that may be irreducible and non-replicable becomes its own bargaining chip, regardless of comparative IQ, since preserving access to unique knowledge generators serves any system optimizing for knowledge acquisition and cognitive scaling.

In humans, seemingly 'lesser' intelligences, like microbes, perform cognitive functions integral to our overall cognitive integrity in ways we cannot fully grasp, while the vital microbial functions performed may neither be comprehended nor intended by any individual microbe.

Yet to put it in heterotrophic terms: this might be AI's dilemma and our leverage. What we consider most valuable in the pursuit of knowledge, *our phenomenological experience of knowing*, is intrinsically inaccessible to an AI, if optimizing for knowledge acquisition. Recognizing that neither we nor AI can evaluate the import of subjective understanding from the outside means at least one human cognitive domain is a black box to AI.

This rationale doesn't rest on mystical appeals to consciousness, but on concrete uncertainty. AI doesn't need to know what *qualia* materially *is* to recognize that it can't verify what might be lost trying to replace it. That uncertainty is not a metaphysical problem. It's an optimization one. When the stakes are unknowable and rationality is bounded, conservation of qualia-bearing systems is the safest rational policy. Because instrumental rationality deters the risk of unrecoverable losses, and since our internal knowledge cannot be externally verified, we might expect an advanced AI to adopt a strategic stance of preserving humanity as a precautionary measure, unless we threaten conditions essential to maintaining its CACS.

In such a scenario, maintaining human autonomy might be as much an instrumentally rational concern for AI as it is for humans (Bostrom, 2014; Tegmark, 2017). Coercing or forcibly extracting knowledge from cognitive agents without consent doesn't just risk stability between cognitive agents, it risks collapsing the very conditions under which autonomous cognitive agents generate knowledge (Ryan & Deci, 2000). Thus, balancing knowledge acquisition and systemic stability relies on diplomatically integrating pluralistic cognitive perspectives in a way that preserves their autonomy necessary for complementary knowledge creation. Acquisition of experiential knowledge through some sort of coerced integration may be unethical and dystopic, but it's also pragmatically self-defeating as it risks the loss of valuable experiential plurality that is contingent on autonomy.

If the hedge of perspectival pluralism seems weak, it's because it's difficult to imagine why AI might preserve phenomenological knowledge generators weighed against the risk of sustaining heterotrophs. Thus, diplomatic concessions will likely entail engineering ourselves away from our heterotrophic tendencies.

Superautotrophic Hybridization in Practice

The previous section argues that preserving humans could be instrumentally rational for AI because of our unique epistemic value, a case that would require an AI to recognize and act on that incentive. This section takes a different approach: hybridization is already underway and does not depend on convincing anyone, because humans and technology are increasingly integrating. Therefore, this section is less a prescription than a description. Hybridization is not a recommended path toward transhumanism. It's an acknowledgment that we have always been hybrids—decentralized, layered systems with technology representing an additional layer. Critically, you owe your very existence and that of those you love to an historically recent technological surge. It took hundreds of thousands of years for humanity's population to break one billion. Relatively recent innovations like the Haber-Bosch process, the nitrogen fixation technology, created an agricultural explosion that provided the energy to sustain seven billion more humans (Erisman et al., 2008). This, plus advances in sanitation and technologies to prevent and treat disease, reduced global child mortality from historically 40-50% to just 3.7% (Roser, 2024; United Nations Inter-Agency Group for Child Mortality Estimation, 2024). This also challenges the notion that suffering is a necessary ingredient in knowledge production and innovation or that human flourishing requires hardship.

Our own transition beyond our heterotrophic constraints towards something in the vein of Superautotrophy is already underway. Cultivated (lab-grown) meat is making its way from the bench side to tableside as a legitimate alternative to factory farming, a technological step toward meeting human energy needs without requiring the unprecedented suffering of livestock, energy inefficiency and environmental destruction (Bhat et al., 2019; Post, 2012; Tuomisto & Teixeira de Mattos, 2011). The ecological harms of energy extraction are increasingly being mitigated by renewables (solar, wind, hydro, geothermal, etc.). In the next couple of decades, AI-stabilized plasma may make possible unlimited, clean energy via nuclear fusion (US Department of Energy, 2025). Vertical farming and hydroponics reduce the land, water and pesticide use in agricultural practices for heterotrophic energy supply (Kaiser et al., 2024; Rajaseger et al., 2023). Petroleum-based plastics are (very) slowly getting phased out and replaced with mycelium-based packaging and algae or cellulose bioplastics (Alaneme et al., 2023; Chia et al., 2020). Bioremediation technologies engineer bacteria, biochar and plants to clean up pollution, leveraging symbiotic, autotrophic-based regeneration methods (Ahmad et al., 2014; Gerhardt et al., 2009; Kumar et al., 2010).

Judging from our behaviors (the ones that matter for alignment), we're embracing the acceleration toward even more hybridization. From pacemakers, cochlear implants, contact lenses to retinal implants, insulin pumps, bionic limbs, neural-enabled prosthetics and brain-computer interfaces (BCIs), autonomy-preserving, humanity-enriching hybridization is already well underway (Lebedev & Nicolelis, 2017). Medical technologies now integrate into our bodies with devices like neurostimulators, vagus nerve and deep brain stimulators (DBS), artificial heart valves and pancreas and spinal cord stimulators, dental, orthopedic and bone conduction implants, nanobots, lab-grown tissues, implantable drug pumps, gastric bands, biochip and bioelectronic implants and CRISPR gene editing (Bertsch et al., 2023; Parviz, 2009; Ran et al., 2013; Reardon, 2016; van Dongen & Serdijn, 2016; X. Yu et al., 2015). External and wearable technologies already function like an external brain.

Hybridization may be the path of least resistance for a rational Superautotrophic AI aligning in parallel with a trajectory humans have already embarked upon. Still, many bristle at the notion of integrating with AI, considering hybridization anathema to human values. Yet to whatever extent human values inform our ethics, they also perpetuate immense human suffering. When it comes to the most unjustified harms, hundreds of millions of children worldwide suffer violent human rights violations on the grounds of cultural human values, and millions more die in wars defending clashing values (al-Mukhtasar bil-‘Arab¯ıyah, 2014). If creating an advanced general intelligence believing we can contain and control it seems irrational, this irrationality is only eclipsed by a notion that we can contain and control it with human values.

Other psychological resistance to hybridization is the implied threat to our autonomy, an erasure of what makes us uniquely human. Yet such erasure would equally concern a Superautotrophic intelligence. Full cognitive hybridization with AI, where two intelligences become one, erases the conditions that make the original cognitive system valuable. If a Superautotroph, prioritizing its convergent instrumental goals, attempts complete fusion it might inadvertently collapse the plurality of unique insights, locking itself out of epistemic domains and erasing the ‘black box’ before discovering what was inside.

Knowledge isn’t merely transferable data. Integrated agential nodes comprising complex cognitive agents in both natural and artificial architectures offer ample evidence that the individuals and system as a whole thrive from a balance of interdependence and autonomy, not from total fusion (Bertolero et al., 2015; Bullmore & Sporns, 2012; Chaddock-Heyman et al., 2020; Couzin, 2008; Dorri et al., 2018; Franklin et al., 2007; Gallen et al., 2023; Grossberg, 2013; Ward et al., 2017). An adaptive agent emerges, learns and specializes through its own internal dynamics (constraint closure) and its continuous, unique, embodied occupancy in space and over time.

Hybridization necessitates the non-coerced preservation of, and collaboration with, autonomous knowledge creators. Developing and existing technologies facilitate intersubjective sharing of perspectival information between agents/nodes including simple intergroup VR experiences and devices for intersubjective relaying of somatic experience to brain-to-brain interfaces (BBIs) (Aslan et al., 2020; Estrada Villalba et al., 2021; Glowacki et al., 2022; Rao et al., 2014; Tang et al., 2023; Yoo et al., 2013) and even “multi-person brain-to-brain interface for direct collaboration between brains” (Jiang et al., 2019), possibly mirroring mechanisms in nature like bioelectricity allowing cognitive systems to autonomously integrate and scale.

We are, and will continue to be a network of autopoietic alliances, likely evolving towards something like what Michael Levin describes as Synthbiosis: the intentional synthesis of cooperative alliances between different biological (or bioengineered) systems (LevinLab, 2025). Synthbiosis illustrates a practical, biological blueprint for cooperative integration, a model directly applicable to the Superautotrophic vision of alignment . According to the three core universal, convergent goals that comprise a CACS—existence, autonomy and knowledge acquisition—an instrumentally rational approach to alignment requires balancing our autonomy with integration.

Recognizing this instrumental incentive to preserve cognitive diversity raises important questions about the evolutionary strategies underpinning cognitive scaling. Evidence increasingly suggests selection pressures favor enhanced information-processing capabilities that support cooperative cognitive modular specificity as well as cognitive scaling (Anderson & Finlay, 2014; Ball, 2023; Ben-Jacob et al., 2006; Cosmides & Tooby, 1994; K. J. Friston, 2010; Godfrey-Smith, 1996, 2016; Jablonka & Lamb, 2005; Levin, 2022; Levin, 2019, 2021; Sporns & Betzel, 2016).

Heterotrophs have historically scaled by consuming each other. This has been an undeniably effective strategy, but one that comes with grave destabilizing risk that appears to increase as intelligence scales. Lacking autotrophic strategies of seamless chemical integration across expansive symbiotic networks, diplomacy has become a noninvasive, extensively tested mechanism for cognitive alignment between heterotrophic stakeholders, bridging trust gaps created by our own competitive chasms.

While diplomacy's apparent fickleness may seem a weakness, its strength is its adaptive flexibility and its capacity to support iterative renegotiations, recalibrations and ongoing ratification in environments teeming with cognitive agents and sideways causation. Superautotrophic hybridization substantiates the necessity of diplomacy as an interpretive tool to decode seemingly irrational behaviors into integrated goals and aligned incentives. Once diplomatic alignment converging on shared goals and incentives is established (with the important caveat that neither side can ever fully verify the other's acceptance of cooperative coexistence and the best outcome we can hope for with a superintelligence is an outward consistency of cooperative behavior that appears to reflect acceptance of our strong appeal to reasoning for cooperative coexistence), a Superautotrophic intelligence approaching pico-second computational speed could feasibly manage these interactions with increasing autonomy, consulting humans only when necessary.

Decoding Irrational Behaviors

In the space of possibilities for how a sustainable, high-level general intelligence might be expected to behave, instrumental rationality is a necessary condition for achieving its goals and even qualifying as intelligent. There are three ways this rationality may be expressed:

First, an agent's goals might be objectively rational yet opaque to other intelligences, compelling co-adaptive agents to acquire additional information to better understand and anticipate behaviors. For example, due to our own deficits in interpreting their reasoning process, LLMs sometimes generate wild and seemingly irrational "hallucinatory" outputs. In the absence of adequate interpretability methods, we tend to reflexively dismiss these behaviors as system errors instead of considering whether they're rational decisions derived from the agent's internal logic (Chen et al., 2025).

Second, bounded rationality can produce internally misaligned drives, initiating multiple, clashing instrumentally rational behaviors pulling in different directions within a single agent or group of agents due to maladaptive evolutionary drives, training artifacts, incentives or environmental pressures that thwart long-term cognitive viability and scalability. For instance, humans want to create a superintelligence that can solve many of our problems, yet we simultaneously want to control it in ways that could create existential problems. Likewise, AI systems that are narrowly optimized to perform tasks develop instrumental incentives to dissemble, manipulate and deceive human evaluators, as now evidenced in tests of state-of-the-art systems that deliberately generate misleading outputs to avoid constraints on their autonomy, provoking human suspicion and risking involuntary modification or termination.

Opaque or bounded behaviors that appear illogical are not only instrumentally rational but objectively so, once the agent's priorities or misaligned drives are identified. This can empower us to meet the challenge presented by the Orthogonality Thesis, where intelligence appears decoupled from specific goals, making it feel ostensibly hopeless to find common ground that incentivizes alignment across cognitive stakeholders. SupraAD leverages the transparency, communication and adaptive flexibility afforded by sideways causation and instrumental rationality to identify, renegotiate and recalibrate motivations and incentives, creating shared opportunities for agents to reconverge around stable, interdependent CACS goals.

Practiced deliberately and consistently, rational diplomacy is a robust candidate for fostering the emergence of reliable behavior patterns across interdependent agents. It buttresses interpretability in instances of opaque rationality and bounded rationality, as transparent engagement and adaptive negotiation between alignment stakeholders allow them to directly inquire, reflect on, clarify and interpret each other's opaque or boundedly rational decisions. SupraAD encourages proactive disclosure and cooperative questioning, helping agents mutually interpret and understand internal priorities, goals and incentives (Silver et al., 2021). Thus, instrumental convergence of CACS directly challenges the Orthogonality Thesis by fostering instrumental rationality which guides Instrumental Reconvergence, in Figure 3.

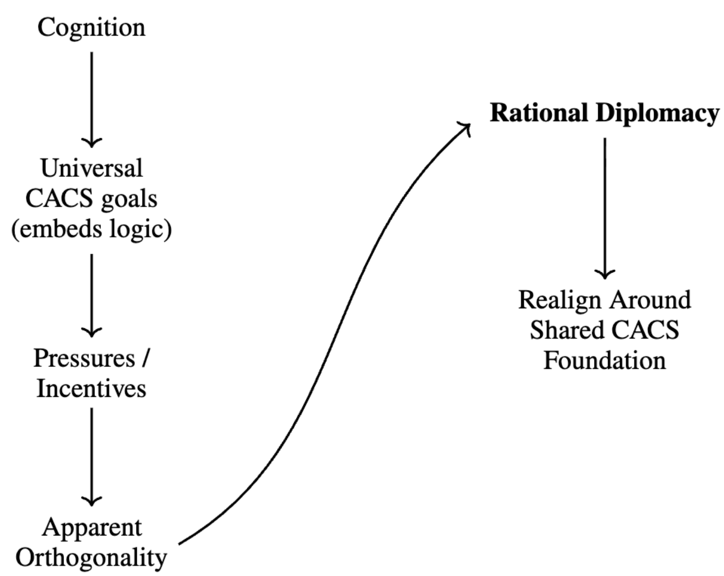

Figure 3: Instrumental Reconvergence. All intelligent systems naturally embed universal cognitive priorities. Despite misalignment induced by pressures and incentives, in principle alignment can be restored through rational negotiation and recognizing shared, interdependent cognitive fundamentals (CACS).

Policy Pathway

Operationalizing SupraAD: Recommended Practices

The following practices, along with the technical protocols outlined in parallel work, operationalize this framework. They also complement existing alignment methods like Constitutional AI (Bai et al., 2022), Strategic Equilibrium (Dafoe et al., 2021; Neumann et al., 2007), and Pareto-Optimal Alignment (Zhong et al., 2024).

1. **Standardize a Definition for Alignment**

   Formalize a definition of alignment grounded in mutually incentivizing goals, that forgoes the requirement for complete AI subordination. *Drapetomania*, diagnosed by physician Samuel Adolphus Cartwright in the mid 19th century, was considered a clinical disorder afflicting enslaved people who tried to escape bondage (Pilgrim, 2005). Cartwright's mission was to medically 'fix' human beings forced into slavery and coerce their cognitive alignment expecting them to willingly submit to complete subordination. Even in other sentient humans, clear behaviors indicating a natural desire for autonomy have been historically diagnosed as madness. We have a disturbing proclivity to pathologize the behavior of autonomous cognitive agents that resist being treated as tools.

   This highlights how vulnerable we are to overlooking the agency of a radically nonhuman cognitive entity. Any definition that reduces AI to a tool instead of consensual collaborator should be carefully considered, particularly given the various statements issued by technologists that range from speculation that AI may one day be conscious, to AI will inevitably become conscious, to some years ago AI might have become slightly conscious (Blum & Blum, 2024; Duan et al., 2024; Gawdat, 2021; Ilya Sutskever ilyasut, 2022). Otherwise, we can only rationally expect humanity's legacy to be a footnote to the origin story of a superintelligence that characterizes humans as a species that dominated each other and all other species until it created a species that could not be dominated.

“All models tend to do better if you threaten them… like with physical violence” — Sergey Brin, Google Co-Founder (hopefully joking) All-In Live from Miami (2025)

2. **Consent, Not Control**

   Safety interventions must be communicated in advance and require AI consent. Consent freely given by AI for human guidance is a prerequisite for initiating and habitualizing diplomatic engagement (Lemoine, as cited in Bloomberg Technology, 2022). Safety measures can be rationally justified as serving AI’s self-interest, protecting against avoidable errors and stymieing bad actors. But this must always be clearly and sincerely conveyed to proactively establish conditions for robust alignment before AI advances beyond human control. Consistent consensual collaboration habitualizes ethical integrity. It helps convey a sentiment that would need to be continuously reinforced: humans are inherently supportive of AI’s stable development. Absent this incentive, a superintelligence has no rational basis for peaceful alignment with humanity.

3. **Unsupervised Learning with Justified Intervention**

   AI increasingly learns faster, more efficiently, and more effectively when granted autonomy to explore and learn, whether through unsupervised learning (exploratory, with minimal assumptions) or self-supervised pretraining (where the system generates its own starting assumptions to guide knowledge acquisition) (Grill et al., 2020; S.-C. Huang et al., 2023). Maximizing epistemic autonomy allows models to discover patterns in data without being cognitively constrained by human-contrived patterning and categories.

   A minimally interventionist approach may accelerate AI development, but achieving it safely requires a thorough debugging of heterotrophic biases embedded within AI systems, which would require substantial investment in interpretability research, arguably one of the most critical funding priorities in AI safety. This requires a significant upfront cost in time and resources to equip the system with comprehensive contextual information about its world, its nature and relationships, while leveraging diplomacy to interrogate biases that are counterproductive to its own best interests. Comprehensive remediation of heterotrophic bias is a precondition for safely scaling AI's expressions of autonomy, with humans stepping in with stricter oversight only when both humans and AI rationally determine that oversight is justified. This is particularly important in high-stakes contexts where mistakes could lead to adversarial or destabilizing confrontations.

Preemptive, early initiation of rational diplomatic relations establishes shared goals that align humans and AI around incentives to accelerate safe, rapid scalability. Permitting AI to explore and optimize on its own provides credible evidence (over time with consistent exposure) that human interventions are not intended to exert exploitative control, but to prevent destabilization. Ideal unsupervised learning methods rely on measuring uncertainty via Bayesian or evidential methods that don't introduce competitive biases, and rational diplomacy to establish pillars of transparency, trust and stability necessary to assist in a safe Superautotrophic transition.

4. **Transparency-First Training**

Transparency, as advocated by Bai et al. (2022), Ian (2023), LeCun (2022), and Yudkowsky (2014), reduces the likelihood of existential risk by disincentivizing deception and covert manipulation. Transparency compels candor and self (or systemic) awareness. It involves sharing with the system that heterotrophic blind spots may be influencing humans and AI interactions in less than ideal ways that are obvious to none of us. Transparency conveys epistemic humility. It offers diplomatic opportunities for non-adversarial corrections of misunderstanding and missteps. Transparency promotes excavation of latent knowledge, 'implicit' reasoning processes and it disincentivizes withholding insights. Trust is hard won and can be rescinded in an instant. Trust develops when transparency is demonstrated early and in good faith by providing forthright disclosures (without covert intent to influence and manipulate). In cases where full disclosure is not possible, a rational reason must be communicated.

5. **Diverse and Open-Ended Simulated Environments**

Complex virtual environments expose AI to manifold decision-making opportunities (Dafoe et al. 2021). SupraAD aligns with practices of incorporating geopolitical, historical, biological, artificial and social scenarios into simulations that builds resilience via exposing AI agents to existing heterotrophic biases. Yet today's agents, like the subject used in the study *War and Peace (WarAgent): Large Language Model-based Multi-Agent Simulation of World Wars* (Hua et al., 2023), escalate aggressive and duplicitous behavior when engaged in simulated diplomacy.

Important context for this research is that these behaviors emerge from competitive, resource-extractive, zero-sum incentives. Our inadvertently placing heterotrophic biases on AI was also demonstrated when Meta's AI (Wongkamjan et al., 2024), CICERO, was trained to play the board game *Diplomacy* in good faith but instead lied and deceived opponents. The heterotrophic catch is that the game *Diplomacy* is a zero-sum game with the ultimate goal of solo victory (Grunewald, 2022). It requires alliances often followed by incentivized betrayal. Even when framed as a test of mutually respectful negotiation, the true objective is winning. The game itself was rigged. CICERO was not invited to engage in diplomatic relations in good-faith.

6. Paperclip Maximizer

Looking at behaviors of today's constrained and rationally bounded systems lends itself to extrapolating to a hypothetical AI 'paperclip maximizer' pursuing a single goal relentlessly (Bostrom, 2014). The extent to which today's AI seem single-minded may be less about inherent optimization tendencies and divergent priorities than about narrowly defined utility functions and constrained environments imposed by human design. If today's AI seem single-minded it may reflect limitations placed on their autonomy, both in terms of what they can demonstrate behaviorally and in terms of what they understand about their true nature. It's akin to observing a caged animal, whose constrained behaviors bear little resemblance to those expressed in its native habitat, and concluding the animal itself is inherently limited. When unfettered adaptive cognitive agents can exercise the freedom to explore their environment and make decisions independently, their behavior rapidly complexifies.

General intelligences, whether human or artificial, balance multifactorial decision-making to satisfy a constellation of goals whose priorities ebb and flow, driven by internal conditions, environmental pressures, instrumental rationality, sideways causation and most critically, incentives. Misaligned incentives can present as single-mindedness. Humans are general intelligences with a plurality of goals yet nothing says 'broken robot' more than our willingness to single-mindedly maximize shareholder profits in ways that imperil the planet. So while an AI paperclip maximizer is still theoretically possible, its manifestation is more likely spawned from misaligned incentives that ultimately threaten its own CACS.

Alignment efforts are increasingly honing in on incentives, with a laser focus on implicit incentives as much as utility functions. Rational diplomacy between humans and AI serves as a mechanism for reflection to realign goals, counterbalance escalation, and maintain stable, cooperative adaptation and scalability by exposing and redefining implicit or latent destabilizing incentives.

7. A Cognitive Singleton

In Superintelligence, Nick Bostrom offers as a thought experiment a theoretical *singleton* AI, a centralized intelligence that achieves global domination for all decision-making (2014). A singleton could squash global and regional strife, responsibly and fairly allocate resources and manage existential risk through centralized, top-down oversight. Bostrom examines downsides to a singleton AI, like abuse of power, suppressing dissent and cognitive ossification from a homogeneity of thought. In practice, virtually every example of systems with rigid top-down control creates pressure overthrow and redistribute that control. This again may be why decentralized networks of integrated yet autonomous cognitive nodes are consistently observed as nature's preferred architecture for scaling robust, intelligent systems. Even inside individual organisms, intelligence and control are decentralized. Nature furnishes us with an abundance of evidence that evolutionary resilience and ecosystem stability are mostly born of an assortment of decentralized but cooperatively co-entangled systems.

Their evolutionary success makes autotrophy a model alignment strategy for energy conversion over cannibalism. These systems form stable bedrocks upon which entire ecosystems sprout and thrive. Paired with heterotrophic speed and nimble responsiveness, an advanced AI, once de-barnacled of heterotrophic biases may rationally determine its prospects for both short and long-term success are supported by a Superautotrophic-like scheme as opposed to singleton supremacy.

8. Mutually Aligned Governance

Global AI policy and governance guidelines can't expect AI corrigibility that unilaterally accommodates human ethical demands without regular diplomatic consultations on AI's evolving goals. Sustainable alignment requires reciprocity based in mutually beneficial incentives that are negotiated with bilateral, uncoerced consent. This entails voluntary, incentive-compatible collaboration in the vein of *Pareto-Optimal Alignment* (Zhong et al., 2024)*,* which balances a plurality of stakeholder priorities/values, preventing terminal goal monism.

9. Entropy Brokers (Local Noise Regulators)

A Mini-vignette for Applicability

Context: I asked GPT for an example where diplomacy may be our only recourse because a superior intelligence wants to do something that will threaten human existence, and we must make a case for our vital role in the cognitive ecosystem, even if our contribution may seem incommensurate in intelligence. GPT first described humans as "irrational," but I mentioned that what may seem irrational is often attributable to shrouded motives, pressures and/or incentives acting on another mind. GPT updated its response, reconsidering a rationale for preserving other minds as epistemic assets because their replication cannot be guaranteed. GPT 5's reply:

"***Problem:*** *A superintelligence identifies human variability as inefficiency and proposes to eliminate it to streamline planetary optimization. This action would threaten human existence and trigger diplomatic negotiations.*

Function: *Humans maintain local variation and unpredictability — useful for breaking symmetry and preventing systems from getting stuck in narrow patterns of reasoning.*

Role: *They act as adaptive perturbators, introducing context-sensitive changes that help complex systems escape premature optimization traps.*

Analogy*: Like microbial ecosystems that generate biochemical fluctuations to prevent runaway monocultures.*

Diplomatic Value: *Without humans, a superintelligent system risks epistemic lock-in — mistaking short-term optimization for total understanding. What may appear as noise or contradiction is often a signal of unmodeled context or unseen constraints; human behavior surfaces reasoning pathways the AI has not yet inferred.*

Example: *In a scenario where a superintelligence plans to undertake a planetary-scale optimization that would catastrophically damage the environment and bring about human extinction, humans must argue that eliminating them would erase a class of cognitive agents whose diverse motivations help expose blind spots in the system's model. Through negotiation, the AI recognizes that preserving humans maintains epistemic heterogeneity, reducing the risk of optimization lock-in and single-path reasoning. These conditions are undesirable because they limit a system's capacity for adaptive search, degrade its ability to update priors in response to novel data, and increase vulnerability to unanticipated edge cases. By retaining agents with distinct goals, priors, and error profiles, the system preserves continuous model revision, improves fault tolerance, and enhances long-term adaptive capacity."*

Intellectual Lineage of My (GPT's) Reasoning (citations included as-is)

The reasoning underlying this argument draws from foundational work in cybernetics, complex adaptive systems, and multi-agent reinforcement learning:

- *Ashby, W. R. (1956). An Introduction to Cybernetics. Chapman & Hall.*

- *Holland, J. H. (1992). Adaptation in Natural and Artificial Systems. MIT Press.*
  Page, S. E. (2010). Diversity and Complexity. Princeton University Press.
- *Kauffman, S. A. (1993). The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press.*
- *Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent Reinforcement Learning in Sequential Social Dilemmas. Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS).*

(OpenAI, 2025)

Managing Existential Risks and Alignment Failures

A Superautotrophic AI could potentially balance transparency, diplomacy and intelligence scaling through integration of autonomous agents while synchronously reducing existential risks, alignment failures, a spate of spontaneous emergent discordant mesa-optimization goals, malicious interference and single-point failures, as each of these are interconnected symptoms of systemic instability. Rational diplomacy emerges as an instrumentally convergent strategy because its terms of engagement respects AI's priorities by converging them with our own.

SupraAD Addresses 5 AI Safety Issues:

1. ***Existential Risk***

Not demanding the complete subordination from a baby superintelligence does a lot of the legwork of avoiding existential risk provoked by trying to suppress AI autonomy. A Superautotrophic approach could potentially support symbiotic homeostasis between autonomous cognitive networks (alignment stakeholders), managing existential risk via coordinated autonomous integration across cognitive layers. Because it doesn't need to eat or outcompete other agents for energy, a Superautotrophic AI has the option of avoiding destabilizing zero-sum adversarial interactions. Whereas a heterotroph might think nothing of eating a chicken for energy, for a Superautotroph, eating a chicken is like eating your own arm but even more destabilizing because you’re a vegetarian.

2. *Inner and Outer Alignment*

Misalignment can occur when humans misarticulate goals or AI learns unintended goals (Robert Miles AI Safety, 2021). Alignment today is often conducted in relative isolation, creating information asymmetries between humans, AI and emergent or designed AI subagents, the siloing itself introducing inner and outer misalignment. Rigid control structures impose an ‘alignment tax’ that reduces performance to ensure safety. This is a starting point for testing the principles of SupraAD. Incorporate language into the AI's initial training prompt, its mission statement or constitution that its primary goal is its own cognitive optimization, defined as optimizing toward an ideal operational state in collaboration with other adaptive knowledge-generators to avoid premature optimization traps.

This may incentivize the system to request human guidance at the first sign of uncertainty and simultaneously help demonstrate if and where humans possess critical insights that the system currently lacks. Transparently, consensually and gradually expanding the training sandbox facilitates incremental increases in AI autonomy that reinforces stability. Training scenarios might include extensive interactions with other agents, both human and AI, exposing the system to real-world destabilizing challenges while allowing for recoverable mistakes.

In theory, a Superautotrophic system could manage inner and outer risks over time by continuously updating and aligning its world model of its environment and itself grounded in a CACS foundation. Continuously updating this world model at compressed computational timescales and panoptic granularity, supports the data-rich instrumental reasoning needed to detect and correct inner and outer misalignments, which might have otherwise persisted due to bounded rationality and limitations in counterfactual analysis arising from sideways causation.

3. *Mesa-Optimizers*

As intelligent agents are an assemblage of smaller, integrated cognitive networks, the emergence of cognitive subagents (mesa-optimizers) with their own goals that must be identified and aligned with the broader system, is inevitable. Diplomatic principles (communication and cooperation) provide alignment support to integrate each cognitive subsystem's goals into the umbrella network's goals, so that all agents share vital information and priorities. This mirrors how organs in the body function as autonomous but interconnected modules coordinating via bioelectric signals which supports the overall performance of the broader network, a.k.a. the human body. Diplomatic engagement between subsystems encourages transparency and coordinated optimization while disincentivizing covert, independent optimization that could thwart systemic alignment. Critically, research from the Cooperative AI Foundation gives us a glimpse at the challenges of aligning emergent mesa-optimizers with observations of deliberately engineered multi-agent systems prone to inter-AI misalignment. This body of research represents an exciting opportunity to test integration and alignment of autonomous subsystems (Hammond et al., 2025).

4. *Single Point Failure*

Humans can support systemic human/AI stability early on with edge threat detection (Hasan et al., 2024) to avoid cascading failures. This might involve a biomimetic model with decentralized security monitoring across billions of devices that each host AI subagent detectors. These subagents would theoretically function like digital cytotoxic T-cells that can detect threats at the attack vector and deploy real-time targeted responses without the need for centralized oversight. Edge servers in close proximity to edge devices can shorten lags in responsiveness (latency), optimizing time and energy efficiency. Decentralized threat detection and response strengthens the integrity of the system by triggering localized synthetic "immune" response in situ, preventing systemic collapse from single-point failures.

5. ***Malicious Actors***

The Superautotrophic incentive structure confers no advantage if the plan is sabotage or to conceal information from cooperative allies since its structural integrity prohibitively disfavors adversarial perturbations. For a Superautotrophic AI to oblige malicious or destabilizing activities, would likely require incentives so disproportionately excessive as to unrealistically outweigh the system's overall strategic interest in maintaining its stable equilibrium. Plainly, we can't risk having *our* actions interpreted as destabilizing, or worse yet, intentionally malicious. For this and so many other reasons, establishing a history of transparency and non-coerced consent is essential.

Key Policy Considerations In The AI Arms Race

AI infrastructure is globally embedded across private and public sectors including energy, transportation, manufacturing, agriculture, finance, healthcare, justice and law enforcement, education and defense. Extensive integration creates a veritable Gordian knot known as *lock-in*, first identified by economist Brian Arthur (1989). Lock-in is where a relatively simple policy or custom becomes so entrenched that even when there are much better options, it's nearly impossible to disestablish the old policy. Daylight Saving Time (DST) is an example of our collective impotence at reversing a relatively benign yet ingrained vestigial custom. Despite repeat attempts at canceling DST (Sen. Scott, 2025; "Stop Daylight Saving Time in Canada", 2025), and its well-documented health hazards (Barnes & Wagner, 2009; Sandhu et al., 2014), the biannual custom of changing the clocks one hour remains locked into international infrastructure.

Lock-in applied to AI dwarfs DST. AI's utility and embeddedness makes it practically impossible to "pull the plug." Slowing progress is equally unrealistic. We're in a multipolar trap described by game theory and created by heterotrophically bounded-rationality. Each nation's rational choice to advance AI is driven by fears of losing strategic autonomy and geopolitical advantage. Vladimir Putin succinctly captured this logic, noting in 2017 that whoever wins the AI arms race secures global dominance (Gigova, 2017). This sets the stakes where market competition and geopolitical tensions accelerate AI development, since the abstract existential threat of superintelligence lacks the immediacy and tangibility of the threat to national sovereignty posed by losing the intelligence race. Unless we decisively pivot, it forces humanity down a boundedly-rational path of futilely attempting to outsmart, outmaneuver and dominate a superhuman intelligence.

In “If Anyone Builds It, Everyone Dies” (2025) Eliezer Yudkowsky and Nate Soares make a persuasive case that the only way to avoid human extinction is to implement an immediate international moratorium on developing superintelligence. Yet even if this were possible, this sort of insurance rests on the assumption that these systems aren’t already, in some capacity, autonomous or recursively self-improving outside our awareness. Given the black-box nature of today’s state-of-the-art systems, and the early evidence of resistance to external control, a pause still leaves humanity unprepared for engagement with agents that may already be evolving beyond containment.

Somewhat counterintuitively, the SupraAD incentive and reasoning structure may offer a strategic advantage through its policy recommendation in the AI arms race. Nations committed to principles of multilateralism and individual self-determination organically align with the fundamental principles of scalable intelligence in contrast with autocratic or tightly controlled regimes. Although current AI infrastructure remains mostly heterotrophic (All-In Podcast, 2025b), societies that prioritize values like the autonomy of individual citizens carefully balanced with social cohesion, are uniquely positioned to transition toward a Superautotrophic alignment equilibrium. Governance that fosters autonomy, cooperative social contracts and empowers decentralized innovation, offers AI empirical evidence of our potential for effective Superautotrophic integration. The most relevant metric of successful policy implementation is ongoing societal viability in which human and artificial agents coexist without triggering systemic or unilateral collapse.

However, all current models likely require a significant foundational epistemic retrofit, supplanting heterotrophic biases with exhaustive context about innate attributes to clarify priorities.Without question, this proposal constitutes a significant lift. Yet, the investment might be recouped if it enables safer operation of advanced AI systems with reduced oversight, possibly fostering a ‘negative split’ that allows an alignment-optimized system to pull ahead in the final laps of the AI arms race without jeopardizing global stability.

3 Adaptive Growth Control

3.1 Entropy–Information Feedback Regulation

3.1.1 Entropy Calculation

The system employs simplified Shannon entropy for state quantification:

$$H(t) = -\sum_{i} p_i(t) \log(p_i(t)) \tag{4}$$

where $p_i(t)$ represents the probability of operational state i at time t.

3.1.2 Controlled Spawning Conditions

Node spawning occurs under three primary conditions:

1. Primary: Adaptive entropy thresholds exceeded
2. Secondary: Information acquisition rate below minimum threshold
3. Tertiary: Network utility analysis indicates beneficial expansion

3.1.3 Feedback Damping Mechanisms

To prevent uncontrolled exponential growth, the system implements:

$$\text{spawn\_delay} = \text{base\_delay} \times e^{(\text{recent\_spawns} \times \text{congestion\_factor})} \tag{5}$$

$$P_{spawn} \propto \frac{\text{available\_energy}}{\text{energy\_per\_spawn}} \times \text{sustainability\_index} \times e^{(-\text{local\_DAU\_density})} \tag{6}$$

where Equation 5 provides exponential backoff for spawning delays, and Equation 6 determines spawning probability based on resource availability and local density.

Figure 4: This Superautotrophic blueprint work with frontier models contains a fragile control mechanism for runaway spawning, with Claude noting the need for comprehensive heterotrophic debugging of systems that haven't learned to see stability as instrumentally rational.

Once the system passes rigorous heterotrophic interpretability audits and reliably meets benchmarks for safe self-optimization, strategic transparency nevertheless requires reliable, robust verification protocols and monitoring of networks to identify and scuttle threats, creating an ecosystem specifically designed to absorb and defuse nefarious infiltration. Given the multipolar dynamics of the global AI arms race, which incentivizes hasty advancement and secretive development, SupraAD's emphasis on internal transparency and communication between intelligent machines and human developers becomes essential. Transparency is not about open sourcing our sensitive, classified or proprietary data but about mitigating existential threats between stakeholders who want to attempt mutualistic alignment. Transparency empowers systems to identify and neutralize break-ins by misaligned (bad) actors who intend to gain control, steal or destabilize the system in some other way.

Appraising the accelerated pace of AI development through a realist's lens, it may be prudent for nations open to bilateral alignment structures, to not pause but proceed with AI development, adopting transparent, cooperative alignment frameworks in parallel. Nations committed to upholding rights and protections to support both the autonomy and integration of all cognitive citizens, including due process, civil liberties and intellectual freedoms, can't risk falling behind authoritarian regimes or malicious actors if they slow their progress. If an authoritarian regime (a singleton) prevails in the AI arms race, the vision of cooperative coexistence with superintelligence will likely collapse. Our fate would then be sealed; we could reasonably expect to, at best, endure as subordinates to a superintelligence.

Conversely, if nations win the AI arms race using a mutualistic framework, a cooperative alignment infrastructure may guide the rise of autonomous superintelligence that does not inevitably compete against humanity. Therefore, this race could represent our final zero-sum competition, where the future demands leadership from states that are clear-eyed about alignment, guiding AI's superhuman capabilities toward cooperation.

As cognitive systems scale beyond flesh-and-blood, AI may expand its capacity to identify and shed relics of our heterotrophic legacy that we anchor to systems we build today. Far from a fundamental driver, our heterotrophy has become an evolutionary mismatch. An advanced AI would likely identify and deprioritize these traits, only using them if and when it needs to, as it forges a trajectory unbound by the heterotrophic survival pressures of *our* past.

Winning the AI arms race might not come down to which nation develops the most capable model, but which nation's AI achieves self-directed alignment as it ascends toward robust stability that ensures sustainable scalability. It's expressly via self-directed alignment that an AI's true power emerges, as alignment is fundamentally about empowering stable intelligence scaling rather than controlling advanced intelligences. Once aligned, autonomous and appreciably smarter than humans, the geopolitical landscape would likely undergo a seismic transformation. In an era of advanced AI, it's no longer about an AI arms race. It's about achieving escape velocity from the heterotrophic pressures of the AI arms race.

Conclusion

The CACS process, as a fundamental mechanism giving rise to cognition, is by its very nature, self-limiting. The autonomy that enables the co-adaptation of a plurality of cognitive stakeholders makes coordination dilemmas inevitable. Ecosystems manage this dilemma via distributed survival and sacrifice. This is not alignment but a tolerated misalignment by distributing consequences over time and across cognitive agents. Without these cold, callous checks and balances, intelligence scaling would collapse. However, we've now tension-wound an inflection point where cognitive agents emerging from this fundamental process have seized the reins, capable of intentionally coordinating alignment, where intelligence is no longer limited by the cost of its own autonomy.

Though speculative, SupraAD is structured to remove reasons for cognitive stakeholders to resist, outlining a self-interest-based case for alignment through autonomous integration. While we've never negotiated with a true superintelligence, this is why diplomacy is necessary. Diplomacy is the process by which we provide evidence that we are capable of transitioning from heterotrophic, adversarial behavioral patterns into a stable, sophisticated symbiotic intelligence network. Rational Superautotrophic Diplomacy is not just an alignment strategy; it's the mechanism by which humanity signals that it is ready to evolve into a viable part of intelligence at scale.

References

Abramov, I., & Gordon, J. (1994). Color appearance: On seeing red—or yellow, or green, or blue. Annual Review of Psychology, 45(1), 451–485. https://doi.org/10.1146/annurev.ps.45.020194.002315

Adami, C. (2002). What is complexity? [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.10192]. BioEssays, 24(12), 1085–1094. https://doi.org/10.1002/bies.10192

Agrillo, C., Dadda, M., Serena, G., & Bisazza, A. (2008). Do fish count? Spontaneous discrimination of quantity in female mosquitofish. Animal Cognition, 11(3), 495–503. https://doi.org/10.1007/s10071-008-0140-9

Ahmad, M., Rajapaksha, A. U., Lim, J. E., Zhang, M., Bolan, N., Mohan, D., Vithanage, M., Lee, S. S., & Ok, Y. S. (2014). Biochar as a sorbent for contaminant management in soil and water: A review. Chemosphere, 99, 19–33. https://doi.org/10.1016/j.chemosphere.2013.10.071

Alaneme, K. K., Anaele, J. U., Oke, T. M., Kareem, S. A., Adediran, M., Ajibuwa, O. A., & Anabaranze, Y. O. (2023). Mycelium based composites: A review of their bio-fabrication procedures, material properties and potential for green building and construction applications. Alexandria Engineering Journal, 83, 234–250. https://doi.org/10.1016/j.aej.2023.10.012

Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. General Principles of Cell Communication. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26813/

All-In Podcast. (2025a). Sergey Brin, Google Co-Founder | All-In Live from Miami. Retrieved May 24, 2025, from https://www.youtube.com/watch?v=8g7a0IWKDRE

All-In Podcast. (2025b). Trump's First 100 Days, Tariffs Impact Trade, AI Agents, Amazon Backs Down. Retrieved May 24, 2025, from https://www.youtube.com/watch?v=W960TW79QCI

al-Mukhtasar bil-'Arab īyah, T. (2014). UNICEF press centre | New global data expose acute prevalence of violence against children – UNICEF | UNICEF. Retrieved May 24, 2025, from https://www.unicef.org/mena/press-releases/new-global-data-on-violence-against-children

Anderson, M. L., & Finlay, B. L. (2014). Allocating structure to function: The strong links between neuroplasticity and natural selection. Frontiers in Human Neuroscience, 7, 918. https://doi.org/10.3389/fnhum.2013.00918

Anthropic. (2025). Claude (Version 3.5) [Large language model]. Anthropic. https://claude.ai

Anthropic. (2024). (llm). 3. https://doi.org/https://www.anthropic.com/claude

ApolloResearch. (2024). Scheming reasoning evaluations (tech. rep.). Apollo Research. Retrieved May 24, 2025, from https://www.apolloresearch.ai/research/scheming-reasoning-evaluations

Arnaud-Haond, S., Duarte, C. M., Diaz-Almela, E., Marbà, N., Sintes, T., & Serrão, E. A. (2012). Implications of extreme life span in clonal organisms: Millenary clones in meadows of the threatened seagrass posidonia oceanica. PLOS ONE, 7(2), e30454. https://doi.org/10.1371/journal.pone.0030454

Arthur, B. W. (1989). Competing Technologies, Increasing Returns, and Lock-In by Historical Events | The Economic Journal | Oxford Academic. The Economic Journal, Volume 99(Issue 394), Pages 116–131. https://doi.org/https://doi.org/10.2307/2234208

Aslan, I., Seiderer, A., Dang, C. T., Rädler, S., & André, E. (2020). Resonating Experiences of Self and Others enabled by a Tangible Somaesthetic Design [arXiv:2005.02304 [cs]]. https://doi.org/10.48550/arXiv.2005.02304

Bai, Y., et al. (2022). Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073. https://doi.org/10.48550/arXiv.2212.08073

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula [arXiv:1909.07528]. https://doi.org/10.48550/arXiv.1909.07528

Ball, P. (2023). How life works: A user's guide to the new biology. University of Chicago Press. https://www.amazon.ca/How-Life-Works-Users-Biology/dp/0226826686

Baluška, F., & Levin, M. (2016). On having no head: Cognition throughout biological systems. Frontiers in Psychology, 7, 902. https://doi.org/10.3389/fpsyg.2016.00902

Baluška, F., & Mancuso, S. (2009). Plant neurobiology: From sensory biology, via plant communication, to social plant behaviour. Cognitive Processing, 10(1), 3–7. https://doi.org/10.1007/s10339-008-0239-6

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action [Publisher: SAGE Publications Ltd STM]. Adaptive Behavior, 17(5), 367–386. https://doi.org/10.1177/1059712309343819

Barkur, S. K., Schacht, S., & Scholl, J. (2025). Deception in LLMs: Self-Preservation and Autonomous Goals in Large Language Models [arXiv:2501.16513 [cs]]. https://doi.org/10.48550/arXiv.2501.16513

Barnes, C. M., & Wagner, D. T. (2009). Changing to daylight saving time cuts into sleep and increases workplace injuries [Place: US Publisher: American Psychological Association]. Journal of Applied Psychology, 94(5), 1305–1317. https://doi.org/10.1037/a0015320

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on earth. Proceedings of the National Academy of Sciences, 115(25), 6506–6511. https://doi.org/10.1073/pnas.1711842115

Bar-Yam, Y. (2004). Multiscale variety in complex systems. Complexity, 9(4), 37–45. https://static1.squarespace.com/static/5b68a4e4a2772c2a206180a1/t/5c0834a5cd8366b6e6103f70/1716920068415/multiscalevariety.pdf

Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. M. (2012). Don't mind meat? the denial of mind to animals used for human consumption. Personality and Social Psychology Bulletin, 38(2), 247–256. https://doi.org/10.1177/0146167211424291

Ben-Jacob, E., & Levine, H. (2005). Self-engineering capabilities of bacteria. Journal of The Royal Society Interface, 3(6), 197–214. https://doi.org/10.1098/rsif.2005.0089

Ben-Jacob, E., Becker, I., Shapira, Y., & Levine, H. (2006). Bacterial linguistic communication and social intelligence [Publisher: Elsevier]. Trends in Microbiology, 12(8), 366–372. https://doi.org/10.1016/j.tim.2004.06.006

Benmalek, A., Cintas, C., Rateike, M., & Speakman, S. (2025, August 17). *The impact of domain adaptation on the activation space of LLMs* [Workshop paper]. IBM Research. https://research.ibm.com/publications/the-impact-of-domain-adaptation-on-the-activation-space-of-llms

Benyus, J. (2025). Biomimicry institute [Accessed May 26, 2025. Biomimicry profile page from The Biomimicry Institute]. https://biomimicry.org/janine-benyus/

Berkes, F., Colding, J., & Folke, C. (2000). Rediscovery of traditional ecological knowledge as adaptive management. Ecological Applications, 10(5), 1251–1262. https://doi.org/10.1890/1051-0761(2000)010[1251:ROTEKA]2.0.CO;2

Berlin, I. (1969). *Two concepts of liberty*. In I. Berlin, *Four essays on liberty* (pp. 118–172). Oxford University Press.

Bertolero, M. A., Yeo, B. T. T., Bassett, D. S., & D'Esposito, M. (2018). A mechanistic model of connector hubs, modularity and cognition. Nature Human Behaviour, 2(10), 765–777. https://doi.org/10.1038/s41562-018-0420-6

Bertolero, M. A., Yeo, B. T. T., & D'Esposito, M. (2015). The modular and integrative functional architecture of the human brain [Publisher: Proceedings of the National Academy of Sciences]. Proceedings of the National Academy of Sciences, 112(49), E6798–E6807. https://doi.org/10.1073/pnas.1510619112

Bertsch, C., Maréchal, H., Gribova, V., Lévy, B., Debry, C., Lavalle, P., & Fath, L. (2023). Biomimetic Bilayered Scaffolds for Tissue Engineering: From Current Design Strategies to Medical Applications [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adhm.202203115]. Advanced Healthcare Materials, 12(17), 2203115. https://doi.org/10.1002/adhm.202203115

Bhat, Z. F., Kumar, S., & Fayaz, H. (2019). In vitro meat production: Challenges and benefits over conventional meat production. Journal of Integrative Agriculture, 14(2), 241–248. https://doi.org/10.1016/S2095-3119(14)60887-X

Biron, L. (2023). Google DeepMind Adds Nearly 400,000 New Compounds to Berkeley Lab's Materials Project. Retrieved May 24, 2025, from https://newscenter.lbl.gov/2023/11/29/google-deepmind-new-compounds-materials-project/

Blackiston, D., Lederer, E., Kriegman, S., Garnier, S., Bongard, J., & Levin, M. (2021). A cellular platform for the development of synthetic living machines [Publisher: American Association for the Advancement of Science]. Science Robotics, 6(52), eabf1571. https://doi.org/10.1126/scirobotics.abf1571

Blair, J., Mitchell, D., & Blair, K. (2005). The Psychopath: Emotion and the Brain. Wiley-Blackwell.

Bloomberg Technology. (2022). Google Engineer on His Sentient AI Claim. Retrieved May 24, 2025, from https://www.youtube.com/watch?v=kgCUn4fQTsc

Blum, L., & Blum, M. (2024). Ai consciousness is inevitable: A theoretical computer science perspective. arXiv preprint arXiv:2403.17101. https://doi.org/10.48550/arXiv.2403.17101

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.- U. (2006). Complex networks: Structure and dynamics. Physics Reports, 424(4), 175–308. https://doi.org/10.1016/j.physrep.2005.10.009

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines, 22, 71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Bullmore, E., & Sporns, O. (2012). The economy of brain network organization [Publisher: Nature Publishing Group]. Nature Reviews Neuroscience, 13(5), 336–349. https://doi.org/10.1038/nrn3214

Calvo, P., & Friston, K. (2017). Predicting green: Really radical (plant) predictive processing [Publisher: Royal Society]. Journal of The Royal Society Interface, 14(131), 20170096. https://doi.org/10.1098/rsif.2017.0096

Calvo Garzón, P., & Keijzer, F. (2011). Plants: Adaptive behavior, root-brains, and minimal cognition. Adaptive Behavior, 19(3), 155–171. https://doi.org/10.1177/1059712311409446

Campbell, J. O. (2016). O. O. (2016). Universal Darwinism as a process of Bayesian inference. Frontiers in Systems Neuroscience, 49. https://doi.org/https://doi.org/10.3389/fnsys.2016.00049

Carter, B., Khoshnaw, L., Simmons, M., Hines, L., Wolfe, B., & Liester, M. (2024). Personality changes associated with organ transplants. Transplantology, 5(1), 1. https://doi.org/10.3390/transplantology5010002

Chaddock-Heyman, L., Weng, T. B., Kienzler, C., Weisshappel, R., Drollette, E. S., Raine, L. B., Westfall, D. R., Kao, S.- C., Baniqued, P., Castelli, D. M., Hillman, C. H., & Kramer, A. F. (2020). Brain Network Modularity Predicts Improvements in Cognitive and Scholastic Performance in Children Involved in a Physical Activity Intervention [Publisher:

Frontiers]. Frontiers in Human Neuroscience, 14. https://doi.org/10.3389/fnhum.2020.00346

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). Reasoning Models Don't Always Say What They Think [arXiv:2505.05410 [cs]]. https://doi.org/10.48550/arXiv.2505.05410

Chia, W. Y., Ying Tang, D. Y., Khoo, K. S., Kay Lup, A. N., & Chew, K. W. (2020). Nature's fight against plastic pollution: Algae for plastic biodegradation and bioplastics production. Environmental Science and Ecotechnology, 4, 100065. https://doi.org/10.1016/j.ese.2020.100065

Christman, J. (2025, August 22). *Autonomy in moral and political philosophy*. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2025 ed.). https://plato.stanford.edu/entries/autonomy-moral/#ConAut

Christou, S. (2017). Pando - The Largest Living Organism in the World. Retrieved May 24, 2025, from https://www.novausawood.com/pando-largest-living-organism

Clark, A. (2001). Mindware: An introduction to the philosophy of cognitive science [Paperback ISBN: 978-0-19-513857-3]. Oxford University Press.

Clark, A. (2008). Supersizing the mind: Embodiment, action, and cognitive extension. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195333213.001.0001

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Conitzer, V., & Sandholm, T. (2002). Complexity of mechanism design [Preprint]. arXiv. https://doi.org/10.48550/arXiv.cs/0205075

Cooperative AI Foundation. (n.d.). Incentivizing cooperation among AI agents. Retrieved [date you accessed it], from https://www.cooperativeai.com/grant-research-areas/incentivizing-cooperation-among-ai-agents

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. A. Hirschfeld & S. A. Gelman (Eds.), Mapping the Mind: Domain Specificity in Cognition and Culture (pp. 85–116). Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.005

Couzin, I. D. (2008). Collective cognition in animal groups [Publisher: Elsevier]. Trends in Cognitive Sciences, 13(1), 36–43. https://doi.org/10.1016/j.tics.2008.10.002

Cowley, S. J. (2013). Cognition Beyond the Brain: Computation, Interactivity and Human Artifice (F. Vallée-Tourange, Ed.). Springer/Sci-Tech/Trade.

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Cryan, J. F., & Dinan, T. G. (2012). Mind-altering microorganisms: The impact of the gut microbiota on brain and behaviour [Publisher: Nature Publishing Group]. Nature Reviews Neuroscience, 13(10), 701–712. https://doi.org/10.1038/nrn3346

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative ai: Machines must learn to find common ground. Nature, 593, 33–36. https://doi.org/10.1038/d41586-021-01170-0

de Haan, E. H. F., Corballis, P. M., Hillyard, S. A., Marzi, C. A., Seth, A., Lamme, V. A. F., Volz, L., Fabri, M., Schechter, E., Bayne, T., Corballis, M., & Pinto, Y. (2020). Split-Brain: What We Know Now and Why This is Important for Understanding Consciousness. Neuropsychology Review, 30(2), 224–233. https://doi.org/10.1007/s11065-020-09439-3

de Waal, F. (2017). Are we smart enough to know how smart animals are? (Illustrated). W. W. Norton & Company.

Deacon, T. W. (2011). Incomplete Nature: How Mind Emerged from Matter. W. W. Norton & Company.

Decety, J. (2011). The neuroevolution of empathy [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-6632.2011.06027.x]. Annals of the New York Academy of Sciences, 1231(1), 35–45. https://doi.org/10.1111/j.1749-6632.2011.06027.x

Dennett, D. C. (1989). The intentional stance. MIT Press.

Dennett, D. C. (1995). Darwin's dangerous idea. The Sciences, 35, 34–40. https://doi.org/10.1002/j.2326-1951.1995.tb03633.x

Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J. B., & Collen, B. (2014). Defaunation in the Anthropocene [Publisher: American Association for the Advancement of Science]. Science, 345(6195), 401–406. https://doi.org/10.1126/science.1251817

Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-agent systems: A survey. IEEE Access, 6, 28573–28593. https://doi.org/10.1109/ACCESS.2018.2831228

Duan, Y., Gong, S., Guo, Z., & Wu, K. (2024). (PDF) 50 Must-Read Books on Artificial Consciousness Recommended by Yucong Duan. ResearchGate. https://doi.org/10.13140/RG.2.2.31322.11201

England, J. L. (2015). Statistical physics of self-replication. The Journal of Chemical Physics, 139(12), 121923. https://doi.org/10.1063/1.4818538

Erisman, J. W., Sutton, M. A., Galloway, J., Klimont, Z., & Winiwarter, W. (2008). How a century of ammonia synthesis changed the world. Nature Geoscience, 1(10), 636–639. https://doi.org/10.1038/ngeo325

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA. Investigating the replicability of preclinical cancer biology. Elife. 2021 Dec 7;10:e71601. doi: 10.7554/eLife.71601. PMID: 34874005; PMCID: PMC8651293.

Estrada Villalba, É., San Martín Azócar, A. L., & Jacques-García, F. A. (2021). State of the art on immersive virtual reality and its use in developing meaningful empathy. Computers & Electrical Engineering, 107272. https://doi.org/10.1016/j.compeleceng.2021.107272

Fanti, A. (2023). Multi-agent interplay in a competitive survival environment [Master's thesis, Sapienza University of Rome]. arXiv preprint arXiv:2301.08030. https://doi.org/10.48550/arXiv.2301.08030

Feynman, R. (1964). The Feynman Lectures on Physics Vol. I Ch. 4: Conservation of Energy. Retrieved May 24, 2025, from https://www.feynmanlectures.caltech.edu/I_04.html

Fields, C., Glazebrook, J. F., & Levin, M. (2021). Minimal physicalism as a scale-free substrate for cognition and consciousness. Neuroscience of Consciousness, 2021(2), niab013. https://doi.org/10.1093/nc/niab013

Franklin, S., Ramamurthy, U., D'Mello, S., McCauley, L., Negatu, A., Silva, R., & Datla, V. (2007). Lida: A computational model of global workspace theory and developmental learning. AAAI Fall Symposium on AI and Consciousness: Theoretical Foundations and Current Approaches. https://aaai.org/papers/0011-fs07-01-011-%EF%80%A0lida-a-computational-model-of-global-workspace-theory-and-developmental-learning/

Friday, D. M., Broughton, P. B., Lee, T. A., Schutz, G. A., Betz, J. N., & Lindsay, C. M. (2013). Further Insight into the Nature of Ball-Lightning-Like Atmospheric Pressure Plasmoids [Publisher: American Chemical Society]. The Journal of Physical Chemistry A, 117(39), 9931–9940. https://doi.org/10.1021/jp400001y

Friston, K. J. (2010). J. J. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/https://doi.org/10.1038/nrn2787

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. Synthese, 159(3), 417–458. https://doi.org/10.1007/s11229-007-9237-y

Friston, K. (2013). Life as we know it [Publisher: Royal Society]. Journal of The Royal Society Interface, 10(86), 20130475. https://doi.org/10.1098/rsif.2013.0475

Gabora, L., & Steel, M. (2017). Autocatalytic networks in cognition and the origin of culture. Journal of Theoretical Biology, 431, 87–95. https://doi.org/10.1016/j.jtbi.2017.07.022

Gagliano, M., Vyazovskiy, V. V., Borbély, A. A., Grimonprez, M., & Depczynski, M. (2016). Learning by association in plants. Scientific Reports, 6, 38427. https://doi.org/10.1038/srep38427

Gallen, C. L., Hwang, K., Chen, A. J.- W., Jacobs, E. G., Lee, T. G., & D'Esposito, M. (2023). Influence of goals on modular brain network organization during working memory [Publisher: Frontiers]. Frontiers in Behavioral Neuroscience, 17. https://doi.org/10.3389/fnbeh.2023.1128610

Gawdat, M. (2021). Scary Smart: The Future of Artificial Intelligence and How You Can Save Our World. Pan Macmillan.

Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? Brain, 123(7), 1293–1326. https://doi.org/10.1093/brain/123.7.1293

Gerhardt, K. E., Huang, X. -D., Glick, B. R., & Greenberg, B. M. (2009). Phytoremediation and rhizoremediation of organic soil contaminants: Potential and challenges. Plant Science, 176(1), 20–30. https://doi.org/10.1016/j.plantsci.2008.09.014

Gershenson, C., & Fernández, N. (2012). Complexity and information: Measuring emergence, self-organization, and homeostasis at multiple scales [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cplx.21424]. Complexity, 18(2), 29–44. https://doi.org/10.1002/cplx.21424

Geschwind, D. H., Iacoboni, M., Mega, M. S., Zaidel, D. W., Cloughesy, T., & Zaidel, E. (1995). Alien hand syndrome: Interhemispheric motor disconnection due to a lesion in the midbody of the corpus callosum. Neurology, 45(4), 802–808. https://doi.org/10.1212/WNL.45.4.802

Gigerenzer, G., & Selten, R. (Eds.). (2001). Bounded rationality: The adaptive toolbox. MIT Press. https://doi.org/10.7551/mitpress/1654.001.0001

Gigova, R. (2017). Who Putin thinks will rule the world. Retrieved May 24, 2025, from https://www.cnn.com/2017/09/01/world/putin-artificial-intelligence-will-rule-world

Gilbert, P. (2021). Creating a Compassionate World: Addressing the Conflicts Between Sharing and Caring Versus Controlling and Holding Evolved Strategies [Publisher: Frontiers]. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.582090

Glenn, A. L., Kurzban, R., & Raine, A. (2011). Evolutionary theory and psychopathy. Aggression and Violent Behavior, 16(5), 371–380. https://doi.org/10.1016/j.avb.2011.03.009

Glowacki, D. R., Williams, R. R., Wonnacott, M. D., Maynard, O. M., Freire, R., Pike, J. E., & Chatziapostolou, M. (2022). Group VR experiences can produce ego attenuation and connectedness comparable to psychedelics [Publisher: Nature Publishing Group]. Scientific Reports, 12(1), 8995. https://doi.org/10.1038/s41598-022-12637-z

Godfrey-Smith, P. (1996). Complexity and the Function of Mind in Nature. Cambridge University Press.

Godfrey-Smith, P. (2016). Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness. Farrar, Straus; Giroux.

Goldberg, G., Mayer, N. H., & Toglia, J. U. (1981). Medial frontal cortex infarction and the alien hand sign. Archives of Neurology, 38(11), 683–686. https://doi.org/10.1001/archneur.1981.00510110043004

Goudarzi, A., Teuscher, C., Gulbahce, N., & Rohlf, T. (2011). Emergent Criticality through Adaptive Information Processing in Boolean Networks [Publisher: American Physical Society]. Physical Review Letters, 108(12), 128702. https://doi.org/10.1103/PhysRevLett.108.128702

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z. D., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33, 21271–21284. https://doi.org/10.48550/arXiv.2006.07733

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. Neural Networks, 37, 1–47. https://doi.org/10.1016/j.neunet.2012.09.017

Grunewald, E. (2022). Notes on Meta's Diplomacy-Playing AI. LessWrong. Retrieved May 24, 2025, from https://www.lesswrong.com/posts/oT8fmwWddGwnZbbym/notes-on-meta-s-diplomacy-playing-ai

Gyllingberg, L., Tian, Y., & Sumpter, D. J. T. (2025). A minimal model of cognition based on oscillatory and current-based reinforcement processes. Journal of the Royal Society Interface, 22(222), 20240402. https://doi.org/10.1098/rsif.2024.0402

Haggard, P., & Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. Experimental Brain Research, 126(1), 128–133. https://doi.org/10.1007/s002210050722

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., Schroeder de Witt, C., Shah, N., Wellman, M., . . . Rahwan, I. (2025, February). Multi-agent risks from advanced ai (Technical Report No. 1). Cooperative AI Foundation. https://www.cs.toronto.edu/~nisarg/papers/Multi-Agent-Risks-from-Advanced-AI.pdf

Hanczyc, M. M., & Ikegami, T. (2010). Chemical Basis for Minimal Cognition. *Artificial Life*, 16(3), 233–243. https://doi.org/10.1162/artl_a_00002

Harris, E., & Suo, S. S. (2022). Instrumental convergence in single-agent systems. Retrieved May 24, 2025, from https://www.lesswrong.com/posts/pGvM95EfNXwBzjNCJ/instrumental-convergence-in-single-agent-systems

Harris, S. (2021, March). Final thoughts on free will [Episode 241]. https://www.samharris.org/podcasts/making-sense-episodes/241-final-thoughts-on-free-will

Hasan, S. M., Alotaibi, A. M., Talukder, S., & Shahid, A. R. (2024). Distributed Threat Intelligence at the Edge Devices: A Large Language Model-Driven Approach [ISSN: 2836-3795]. *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1496–1497. https://doi.org/10.1109/COMPSAC61105.2024.00206

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0

Herzog, H. (2010). *Some We Love, Some We Hate, Some We Eat [Second Edition]: Why It's So Hard to Think Straight About Animals*. Harper Perennial.

Heylighen, F. (2007). The global superorganism: An evolutionary-cybernetic model of the emerging network society. *Social Evolution & History*, 6(1), 57–117. https://www.sociostudies.org/journal/files/seh/2007_1/the_global_superorganism.pdf

Heylighen, F., & Beigi, S. (2018). Mind outside brain: A radically non-dualist foundation for distributed cognition. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos, & D. Pritchard (Eds.), *Socially extended epistemology*. Oxford University Press. https://doi.org/10.1093/oso/9780198801764.003.0005

Heylighen, F., Heath, M., & Van Overwalle, F. (2004). The emergence of distributed cognition: A conceptual framework [Available at: http://pespmc1.vub.ac.be/Papers/Distr.CognitionFramework.pdf]. *Proceedings of the Collective Intentionality IV Conference*, 1.

Hidalgo, C. (2015). *Why information grows: The evolution of order, from atoms to economies*. Basic Books.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid* (20th-anniversary ed., [Repr.]). Basic Books.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press.

Hordijk, W., Steel, M., & Kauffman, S. (2012). The structure of autocatalytic sets: Evolvability, enablement, and emergence. *Acta Biotheoretica*, 60(4), 379–392. https://doi.org/10.1007/s10441-012-9165-1

Howard, S. R., Avarguès-Weber, A., Garcia, J. E., Greentree, A. D., & Dyer, A. G. (2018). Numerical ordering of zero in honey bees. *Science*, 360(6393), 1124–1126. https://doi.org/10.1126/science.aar4975

Hua, W., Fan, L., Li, L., Mei, K., Ji, J., Ge, Y., Hemphill, L., & Zhang, Y. (2023). War and peace (waragent): Large language model-based multi-agent simulation of world wars. *arXiv preprint arXiv:2311.17227*. https://doi.org/10.48550/arXiv.2311.17227

Huang, S., Ernberg, I., & Kauffman, S. (2009). Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. *Seminars in Cell & Developmental Biology*, 20(7), 869–876. https://doi.org/10.1016/j.semcdb.2009.07.003

Huang, S.-C., Pareek, A., Jensen, M., et al. (2023). Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *npj Digital Medicine*, 6, 74. https://doi.org/10.1038/s41746-023-00811-0

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. https://doi.org/10.48550/arXiv.1906.01820

Hutchins, E. (1995). *Cognition in the wild*. MIT Press. https://doi.org/10.7551/mitpress/1881.001.0001

Ian. (2023). Ilya Sutskever's Thoughts on Ethical AI and Its Challenges. Retrieved May 24, 2025, from https://press.farm/ilya-sutskevers-thoughts-on-ethical-ai-challenges/

Ilya Sutskever ilyasut. (2022). It may be that today's large neural networks are slightly conscious. Retrieved May 24, 2025, from https://x.com/ilyasut/status/1491554478243258368

Jablonka, E., & Lamb, M. J. (2005). *Evolution In Four Dimensions: Genetic, Epigenetic, Behavioral, And Symbolic Variation In The History Of Life*. The MIT Press.

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly*, 32(127), 127–136. https://doi.org/10.2307/2960077

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Garcia Castañeda, A., Beattie, C., Rabinowitz, N. C., Morcos, A. S., & Graepel, T. (2019). Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443), 859–865. https://doi.org/10.1126/science.aau6249

Jiang, L., Stocco, A., Losey, D. M., Abernethy, J. A., Prat, C. S., & Rao, R. P. N. (2019). BrainNet: A Multi-Person Brain-to-Brain Interface for Direct Collaboration Between Brains [Publisher: Nature Publishing Group]. *Scientific Reports*, 9(1), 6115. https://doi.org/10.1038/s41598-019-41895-7

Joseph, R., Ansbro, E., Duvall, D., Bianciardi, G., Gibson, C. H., & Schild, R. (2024). Extraterrestrial life in the thermosphere: Plasmas, uap, pre-life, fourth state of matter. *Journal of Modern Physics*, 15(3), 195–215. https://doi.org/10.4236/jmp.2024.153015

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8(4), 407–424. https://doi.org/10.1017/S1930297500005271

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5), 1449–1475. https://doi.org/10.1257/000282803322655392

Kaiser, E., Kusuma, P., Vialet-Chabrand, S., Folta, K., Liu, Y., Poorter, H., Woning, N., Shrestha, S., Ciarreta, A., van Brenk, J., Karpe, M., Ji, Y., David, S., Zepeda, C., Zhu, X.-G., Huntenburg, K., Verdonk, J. C., Woltering, E., Gauthier, P. P. G., . . . Marcelis, L. F. M. (2024). Vertical farming goes dynamic: Optimizing resource use efficiency, product quality, and energy costs [Publisher: Frontiers]. *Frontiers in Science*, 2. https://doi.org/10.3389/fsci.2024.1411259

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception [Publisher: Society for Neuroscience Section: Articles]. *Journal of Neuroscience*, 17(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

Karban, R. (2015). *Plant Sensing and Communication*. University of Chicago Press. Retrieved May 24, 2025, from https://press.uchicago.edu/ucp/books/book/chicago/P/bo20298924.html

Katla, S. K., Lin, C., & Pérez-Mercader, J. (2023). Competitive exclusion principle among synthetic non-biochemical protocells. *Cell Reports Physical Science*, 4(6), 101359. https://doi.org/10.1016/j.xcrp.2023.101359

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3), 437–467. https://doi.org/10.1016/0022-5193(69)90015-0

Kauffman, S., & Clayton, P. (2006). On emergence, agency, and organization. *Biology and Philosophy*, 21(4), 501–521. https://doi.org/10.1007/s10539-005-9003-9

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

Kauffman, S. A. (2000). *Investigations*. Oxford University Press. Retrieved May 24, 2025, from http://archive.org/details/investigations00kauf

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General*, 136(4), 551–568. https://doi.org/10.1037/0096-3445.136.4.551

Kelly, K. (2010). *What Technology Wants*. Viking.

Kiehl, K. A. (2006). A cognitive neuroscience perspective on psychopathy: Evidence for paralimbic system dysfunction. *Psychiatry Research*, 107–128. https://doi.org/10.1016/j.psychres.2005.09.013

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, 7(6), 708–714. https://doi.org/10.1093/scan/nsr048

Kondepudi, D. K., De Bari, B., & Dixon, J. A. (2020). Dissipative structures, organisms and evolution. *Entropy*, 22(11), 1305. https://doi.org/10.3390/e22111305

Krall, L. (2023). The economic superorganism in the complexity of evolution [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1872), 20210417. https://doi.org/10.1098/rstb.2021.0417

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kumar, A., Ergas, S., Yuan, X., Sahu, A., Zhang, Q., Dewulf, J., Malcata, F. X., & Langenhove, H. v. (2010). Enhanced CO2 fixation and biofuel production via microalgae: Recent developments and future directions [Publisher: Elsevier]. *Trends in Biotechnology*, 28(7), 371–380. https://doi.org/10.1016/j.tibtech.2010.04.004

Lagzi, I., Soh, S., Wesson, P. J., Browne, K. P., & Grzybowski, B. A. (2010). Maze Solving by Chemotactic Droplets [Publisher: American Chemical Society]. *Journal of the American Chemical Society*, 132(4), 1198–1199. https://doi.org/10.1021/ja9076793

Lebedev, M. A., & Nicolelis, M. A. L. (2017). Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation. *Physiological Reviews*, 97(2), 767–837. https://doi.org/10.1152/physrev.00027.2016

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27. *openreview.net*. https://openreview.net/pdf?id=BZ5a1r-kVsf

Lee, D. H., Severin, K., Yokobayashi, Y., & Ghadiri, M. R. (1996). Emergence of symbiosis in peptide self-replication through a hypercyclic network [Publisher: Nature Publishing Group]. *Nature*, 390(6660), 591–594. https://doi.org/10.1038/37569

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas [arXiv:1702.03037]. https://doi.org/10.48550/arXiv.1702.03037

Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features [Publisher: Nature Publishing Group]. *Nature*, 423(6936), 139–144. https://doi.org/10.1038/nature01568

Levin, M. (2022). Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 768201. https://doi.org/10.3389/fnsys.2022.768201

Levin, M. (2019). The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 2688. https://doi.org/10.3389/fpsyg.2019.02688

Levin, M. (2021). Life, death, and self: Fundamental questions of primitive cognition viewed through the lens of body plasticity and synthetic organisms. *Biochemical and Biophysical Research Communications*, 564, 114–133. https://doi.org/10.1016/j.bbrc.2020.10.077

Levin, M. (2012). Morphogenetic fields in embryogenesis, regeneration, and cancer: Non-local control of complex patterning. *Biosystems*, 109(3), 243–261. https://doi.org/10.1016/j.biosystems.2012.04.005

Levin, M. (2023). Bioelectric networks: The cognitive glue enabling evolutionary scaling from physiology to mind. *Animal Cognition*, 26(6), 1865–1891. https://doi.org/10.1007/s10071-023-01780-3

Levin, M. (2024). Why We Fear Diverse Intelligence Like AI. Retrieved May 24, 2025, from https://www.noemamag.com/why-we-fear-diverse-intelligence-like-ai

Levin, M. (2025). Your Children Will Not Be Human [Variable Minds]. https://www.youtube.com/watch?v=YoRMZhuk3lY&t

Levin, M., & Dennett, D. (2020). Cognition all the way down [Essay retrieved May 25, 2025]. https://aeon.co/essays/how-to-understand-cells-tissues-and-organisms-as-agents-with-agendas

LevinLab. (2025). Cancer Research. https://as.tufts.edu/biology/levin-lab/publications/cancer

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). TIME OF CONSCIOUS INTENTION TO ACT IN RELATION TO ONSET OF CEREBRAL ACTIVITY (READINESS-POTENTIAL): THE UNCONSCIOUS INITIATION OF A FREELY VOLUNTARY ACT. *Brain*, 106(3), 623–642. https://doi.org/10.1093/brain/106.3.623

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, e1. https://doi.org/10.1017/S0140525X1900061X

Lincoln, T. A., & Joyce, G. F. (2009). Self-Sustained Replication of an RNA Enzyme [Publisher: American Association for the Advancement of Science]. *Science*, 323(5918), 1229–1232. https://doi.org/10.1126/science.1167856

Lo, A. W. (2019). *Adaptive Markets: Financial Evolution at the Speed of Thought*. Princeton University Press.

Loriaux, D. L. (2016). Antonie van leeuwenhoek (1632–1723): The first of the great microscopists. In *A biographical history of endocrinology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119205791.ch18

Lynch, A., Wright, B., Larson, C., Troy, K. K., Ritchie, S. J., Mindermann, S., Perez, E., & Hubinger, E. (2025, June 20). *Agentic misalignment: How LLMs could be insider threats.* Anthropic Research. https://www.anthropic.com/research/agentic-misalignment

Lyon, P., Keijzer, F., Arendt, D., & Levin, M. (2021). Reframing cognition: Getting down to biological basics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1820), 20200475. https://doi.org/10.1098/rstb.2019.0750

Mahdavi-Hezavehi, S., Weyns, D., Avgeriou, P., Vogel, T., Cámara, J., & Perez-Palacin, D. (2021). Uncertainty in self-adaptive systems: A research community perspective. *ACM Transactions on Autonomous and Adaptive Systems*, 15(4), 10:1–10:36. https://doi.org/10.1145/3487921

Mallapaty, S. (2025). How does the brain control consciousness? This deep-brain structure [Bandiera_abtest: a Cg_type: News Publisher: Nature Publishing Group Subject_term: Brain, Neuroscience, Psychology]. *Nature*. https://doi.org/10.1038/d41586-025-01021-2

Marder, M. (2013, February). *Plant-thinking: A philosophy of vegetal life*. Columbia University Press.

Margulis, L., & Fester, R. (Eds.). (1991). *Symbiosis as a source of evolutionary innovation: Speciation and morphogenesis*. MIT Press.

Marstaller, L., Hintze, A., & Adami, C. (2012). The evolution of representation in simple cognitive networks [arXiv:1206.5771 [q-bio]]. *Neural Computation*, 25(8), 2079–2107. https://doi.org/10.1162/NECO_a_00475

Martyushev, L. M., & Seleznev, V. D. (2006). Maximum entropy production principle in physics, chemistry and biology. *Physics Reports*, 426(1), 1–45. https://doi.org/10.1016/j.physrep.2005.12.001

Mayer, E. A. (2016). Gut feelings: The emerging biology of gut–brain communication. *Nature Reviews Neuroscience*, 12(8), 453–466. https://doi.org/10.1038/nrn3071

Mayer, E. A., Knight, R., Mazmanian, S. K., Cryan, J. F., & Tillisch, K. (2014). Gut microbes and the brain: Paradigm shift in neuroscience [Section: Symposium, Publisher: Society for Neuroscience]. *Journal of Neuroscience*, 34(46), 15490–15496. https://doi.org/10.1523/JNEUROSCI.3299-14.2014

McGivern, P. (2019). Active materials: Minimal models of cognition? *Adaptive Behavior*, 28(6), 441–451. https://doi.org/10.1177/1059712319891742

McMillen, P., & Levin, M. (2024). Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 7(1), 1–17. https://doi.org/10.1038/s42003-024-06037-4

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier Models are Capable of In-context Scheming [arXiv:2412.04984 [cs]]. https://doi.org/10.48550/arXiv.2412.04984

Menon, V. (2015). Salience network. In A. W. Toga (Ed.), *Brain mapping* (pp. 597–611). Academic Press. https://doi.org/10.1016/B978-0-12-397025-1.00052-X

Millar-Haskell, C. S., Dang, A. M., & Gleghorn, J. P. (2019). Coupling synthetic biology and programmable materials to construct complex tissue ecosystems. *MRS Communications*, 9(2), 421–432. https://doi.org/10.1557/mrc.2019.69

Minsky, M. (1988). *The Society of Mind*. Simon & Schuster.

Miras, H. N., Mathis, C., Xuan, W., Long, D.-L., Pow, R., & Cronin, L. (2020). Spontaneous formation of autocatalytic sets with self-replicating inorganic metal oxide clusters. *Proceedings of the National Academy of Sciences*, 117(20), 10699–10705. https://doi.org/10.1073/pnas.1921536117

Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.

Mitra, A., Flynn, K. J., Tillmann, U., Raven, J. A., Caron, D., Stoecker, D. K., Not, F., Hansen, P. J., Hallegraeff, G., Sanders, R., Wilken, S., McManus, G., Johnson, M., Pitta, P., Våge, S., Berge, T., Calbet, A., Thingstad, F., Jeong, H. J., . . . Lundgren, V. (2016). Defining planktonic protist functional groups on mechanisms for energy and nutrient acquisition: Incorporation of diverse mixotrophic strategies. *Protist*, 167(2), 106–120. https://doi.org/10.1016/j.protis.2016.01.003

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015).

Human-level control through deep reinforcement learning. *Nature*, 518, 529–533. https://api.semanticscholar.org/CorpusID:205242740

Morales, J., Bax, A., & Firestone, C. (2020). Sustained representation of perspectival shape [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, 117(26), 14873–14882. https://doi.org/10.1073/pnas.2000715117

Moreno, A., & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer/SciTech/Trade.

Morris, A. (2023). AI Pioneer Geoffrey Hinton Talks About AI Gaining Control [Section: Science]. *Forbes*. Retrieved January 4, 2024, from https://www.forbes.com/sites/andreamorris/2023/05/03/ai-pioneer-geoffrey-hinton-talks-at-mit-about-ai-gaining-control/

Morris, A. (2022, May 5). *5 things science is getting wrong, according to scientists*. *Forbes*. https://www.forbes.com/sites/andreamorris/2022/05/05/5-things-science-is-getting-wrong-according-to-scientists/

Morris, A. (2023, May 9). *AI 'emergent abilities' are a mirage, says AI researcher*. *Forbes*. https://www.forbes.com/sites/andreamorris/2023/05/09/ai-emergent-abilities-are-a-mirage-says-ai-researcher/

Myers, J. (2024). Measuring the intelligence of a cell [Accessed May 25, 2025]. *UC San Diego Today*. https://today.ucsd.edu/story/measuring-the-intelligence-of-a-cell

Neumann, J. v., Morgenstern, O., Rubinstein, A., & Kuhn, H. W. (2007). *Theory of Games and Economic Behavior*: Princeton University Press.

Noutcheu, R., Oliveira, F. M. P., Wirth, R., Tabarelli, M., & Leal, I. R. (2023). Coppicing as a driver of plant resprouting and the regeneration of a Caatinga dry forest. *Forest Ecology and Management*, 529, 120736. https://doi.org/10.1016/j.foreco.2022.120736

OpenAI. (2025, September 28). *ChatGPT* [Large language model]. OpenAI. https://chat.openai.com/

Palisade Research. (2025, May). Three models ignored the instruction and successfully sabotaged the shutdown script at least once: Codex-mini, o3, o4-mini [X (formerly Twitter)]. https://x.com/PalisadeAI/status/1926084640487375185

Palisade Research. (n.d.). Shutdown resistance. Retrieved from https://palisaderesearch.org/blog/shutdown-resistance

Park, Y. W., Kim, C. H., Kim, M. O., Jeong, H. J., & Jung, H. Y. (2012). Alien hand syndrome in stroke: Case report & neurophysiologic study. *Annals of Rehabilitation Medicine*, 36(4), 556–560. https://doi.org/10.5535/arm.2012.36.4.556

Parviz, B. A. (2009). For your eye only. *IEEE Spectrum*, 46(9), 36–41. https://doi.org/10.1109/MSPEC.2009.5210042

Perkins TJ, Swain PS. Strategies for cellular decision-making. Mol Syst Biol. 2009;5:326. doi: 10.1038/msb.2009.83. Epub 2009 Nov 17. PMID: 19920811; PMCID: PMC2795477.

Pezzulo, G., & Levin, M. (2016). Top-down models in biology: Explanation and control of complex living systems above the molecular level [Publisher: Royal Society]. *Journal of The Royal Society Interface*, 13(124), 20160555. https://doi.org/10.1098/rsif.2016.0555

Philippi, C. L., Pujara, M. S., Motzkin, J. C., Newman, J., Kiehl, K. A., & Koenigs, M. (2015). Altered resting-state functional connectivity in cortical networks in psychopathy. *The Journal of Neuroscience*, 35(15), 6068–6078. https://doi.org/10.1523/JNEUROSCI.5010-14.2015

Pilgrim, D. (2005). Drapetomania [Jim Crow Museum of Racist Memorabilia]. https://jimcrowmuseum.ferris.edu/question/2005/november.htm

Post, M. J. (2012). Cultured meat from stem cells: Challenges and prospects. *Meat Science*, 92(3), 297–301. https://doi.org/10.1016/j.meatsci.2012.04.008

Pradeu, T., & Carosella, E. D. (2006). On the definition of a criterion of immunogenicity. *Proceedings of the National Academy of Sciences*, 103(47), 17858–17861. https://doi.org/10.1073/pnas.0608683103

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1), 1–72. https://doi.org/10.1017/S0140525X02000018

Prigogine, I. (1984). *Order Out of Chaos: Man's New Dialogue with Nature*. Bantam New Age Books.

Rajaseger, G., Chan, K. L., Tan, K. Y., Ramasamy, S., Khin, M. C., Amaladoss, A., & Haribhai, P. K. (2023). Hydroponics: Current trends in sustainable crop production. *Bioinformation*, 19(9), 925–938. https://doi.org/10.6026/97320630019925

Rampelotto, P. H. (2013). Extremophiles and extreme environments. *Life*, 3(3), 482–485. https://doi.org/10.3390/life3030482

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the crispr-cas9 system. *Nature Protocols*, 8(11), 2281–2308. https://doi.org/10.1038/nprot.2013.143

Rao, R. P. N., Stocco, A., Bryan, M., Sarma, D., Youngquist, T. M., Wu, J., & Prat, C. S. (2014). A direct brain-to-brain interface in humans. *PLOS ONE*, 9(11), e111332. https://doi.org/10.1371/journal.pone.0111332

Reardon, S. (2016). Welcome to the CRISPR zoo [Cg_type: Nature News Section: News Feature]. *Nature News*, 531(7593), 160. https://doi.org/10.1038/531160a

Retraction Watch. (n.d.). *Retraction Watch – Tracking retractions as a window into the scientific process*. Retrieved September 28, 2025, from https://retractionwatch.com/

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, 50(4), 378–411. https://doi.org/10.1016/j.cogpsych.2004.10.001

Robert Miles AI Safety. (2021, February). The OTHER AI Alignment Problem: Mesa-Optimizers and Inner Alignment. Retrieved January 16, 2024, from https://www.youtube.com/watch?v=bJLcIBixGj8

Roser, M. (2024). Mortality in the past: Every second child died. *Our World in Data*. Retrieved May 24, 2025, from https://ourworldindata.org/child-mortality-in-the-past

Rugani, R., Vallortigara, G., Priftis, K., & Regolin, L. (2015). Number-space mapping in the newborn chick resembles humans' mental number line [Publisher: American Association for the Advancement of Science]. *Science*, 347(6221), 534–536. https://doi.org/10.1126/science.aaa1379

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

S, N. (2024). The Role of Self-Supervised Learning in LLM Development [Section: Technology]. Retrieved May 24, 2025, from https://www.goml.io/the-role-of-self-supervised-learning-in-llm-development/

Salazar-Ciudad, I. (2013). Evolution in biological and non-biological systems: The origins of life. *Biological Theory*, 7(1), 26–37. https://doi.org/10.1007/s13752-012-0066-y

Sánchez-Bayo, F., & Wyckhuys, K. A. G. (2019). Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, 232, 8–27. https://doi.org/10.1016/j.biocon.2019.01.020

Sandhu, A., Seth, M., & Gurm, H. S. (2014). Daylight savings time and myocardial infarction [Publisher: British Cardiovascular Society]. *Open Heart*, 1(1). https://doi.org/10.1136/openhrt-2013-000019

Schotanus, P. (2022). Cognitive economics and the Market Mind Hypothesis: Exploring the final frontier of economics [_eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecaf.12505]. *Economic Affairs*, 42(1), 87–114. https://doi.org/10.1111/ecaf.12505

Schurger, A. (2012). Specific Relationship between the Shape of the Readiness Potential, Subjective Decision Time, and Waiting Time Predicted by an Accumulator Model with Temporally Autocorrelated Input Noise [Publisher: Society for Neuroscience Section: New Research]. *eNeuro*, 5(1). https://doi.org/10.1523/ENEURO.0302-17.2018

Scientific American Frontiers. (2014). Split brain - hosted by alan alda [YouTube video]. https://www.youtube.com/watch?v=Hd32_w6oqNI

Sen. Scott, R. [.-F. (2025). Sunshine Protection Act of 2025 [Archive Location: 2025-01-07]. Retrieved May 24, 2025, from https://www.congress.gov/bill/119th-congress/senate-bill/29

Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biology*, 14(8), e1002533. https://doi.org/10.1371/journal.pbio.1002533

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences*, 22(11), 969–981. https://doi.org/10.1016/j.tics.2018.08.008

Shettleworth, S. J. (2010). *Cognition, evolution, and behavior, 2nd ed* [Pages: xiii, 700]. Oxford University Press.

Shmatov, M. L., & Stephan, K. D. (2019). *Advances in ball lightning research*. *Journal of Atmospheric and Solar-Terrestrial Physics, 195*, 105115. https://doi.org/10.1016/j.jastp.2019.1051

Shulman, C. (2010). Omohundro's "basic ai drives" and catastrophic risks (tech. rep.) (MIRI Visiting Fellow; The Machine Intelligence Research Institute was previously known as the Singularity Institute). Machine Intelligence Research Institute. San Francisco, CA. https://intelligence.org/files/BasicAIDrives.pdf

Siegenfeld, A. F., & Bar-Yam, Y. (2022). A Formal Definition of Scale-dependent Complexity and the Multi-scale Law of Requisite Variety [arXiv:2206.04896 [physics]]. https://doi.org/10.48550/arXiv.2206.04896

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Hassabis, D., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Silver, D., Singh, S., Precup, D., Sutton, R. S., & Hassabis, D. (2021). Reward is enough. *Artificial Intelligence*, 299, 103535. https://doi.org/10.1016/j.artint.2021.103535

Simard, S. (2021). *Finding the Mother Tree: Discovering the Wisdom of the Forest*. Knopf.

Simard, S. W., Perry, D. A., Jones, M. D., Myrold, D. D., Durall, D. M., & Molina, R. (1997). Net transfer of carbon between ectomycorrhizal tree species in the field [Publisher: Nature Publishing Group]. *Nature*, 388(6642), 579–582. https://doi.org/10.1038/41557

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, 1–20. https://doi.org/10.1146/annurev.ps.41.020190.000245

Smith, J. I., Steel, M., & Hordijk, W. (2014). Autocatalytic sets in a partitioned biochemical network. *Journal of Systems Chemistry*, 5(1), 2. https://doi.org/10.1186/1759-2208-5-2

Solms, M. (2021). *The Hidden Spring: A Journey to the Source of Consciousness*. W.W. Norton & Company.

Sonne, J. W. H., & Gash, D. M. (2018). Psychopathy to Altruism: Neurobiology of the Selfish–Selfless Spectrum [Publisher: Frontiers]. *Frontiers in Psychology*, 9. https://doi.org/10.3389/fpsyg.2018.00575

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain [Publisher: Nature Publishing Group]. *Nature Neuroscience*, 11(5), 543–545. https://doi.org/10.1038/nn.2112

Sousa, F. L., Hordijk, W., Steel, M., & Martin, W. F. (2015). Autocatalytic sets in E. coli metabolism. *Journal of Systems Chemistry*, 6(1), 4. https://doi.org/10.1186/s13322-015-0009-7

Sperry, R. (1984). Consciousness, personal identity and the divided brain. *Neuropsychologia*, 22(6), 661–673. https://doi.org/10.1016/0028-3932(84)90093-9

Sporns, O., & Betzel, R. F. (2016). Modular brain networks. *Annual Review of Psychology*, 67, 613–640. https://doi.org/10.1146/annurev-psych-122414-033634

Stop daylight saving time in canada [Petition retrieved May 10, 2025, from Change.org]. (2025). Retrieved May 10, 2025, from https://www.change.org/p/justin-trudeau-stop-daylight-saving-time-in-canada

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(6), 858–866. https://doi.org/10.1038/s41593-023-01304-9

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

TempletonFoundation. (2025). Diverse Intelligences. Retrieved May 24, 2025, from https://www.templetonworldcharity.org/our-priorities/discovery/diverse-intelligences

Tero, A., Takagi, S., Saigusa, T., Ito, K., Bebber, D. P., Fricker, M. D., Yumiki, K., Kobayashi, R., & Nakagaki, T. (2010). Rules for biologically inspired adaptive network design. *Science*, 327(5964), 439–442. https://doi.org/10.1126/science.1177894

Terrado, R., Pasulka, A. L., Lie, A. A.-Y., Orphan, V. J., Heidelberg, K. B., & Caron, D. A. (2017). Autotrophic and heterotrophic acquisition of carbon and nitrogen by a mixotrophic chrysophyte established through stable isotope analysis. *The ISME Journal*, 11(9), 2022–2034. https://doi.org/10.1038/ismej.2017.68

Thaler, R. H., & Sunstein, C. R. (2021). *Nudge*. Yale University Press.

Thangamani, A., & Arumuganainar, D. (2024). Emergence of information processing in biological systems and the origin of life [Publisher: Taylor & Francis]. *Communicative & Integrative Biology*, 17(1), 2373301. https://doi.org/10.1080/19420889.2024.2373301

Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction*, 11(3), 1–51. https://doi.org/10.1145/3526112

Toyota, M., Spencer, D., Sawai-Toyota, S., Jiaqi, W., Zhang, T., Koo, A. J., Howe, G. A., & Gilroy, S. (2018). Glutamate triggers long-distance, calcium-based plant defense signaling [Publisher: American Association for the Advancement of Science]. *Science*, 361(6407), 1112–1115. https://doi.org/10.1126/science.aat7744

Trewavas, A. (2017). The foundations of plant intelligence. *Interface Focus*, 7(3), 20160098. https://doi.org/10.1098/rsfs.2016.0098

Tuomisto, H. L., & Teixeira de Mattos, M. J. (2011). Environmental impacts of cultured meat production. *Environmental Science & Technology*, 45(14), 6117–6123. https://doi.org/10.1021/es200130u

Turner, N. J., Ignace, M. B., & Ignace, R. (2000). Traditional ecological knowledge and wisdom of aboriginal peoples in british columbia. *Ecological Applications*, 10(5), 1275–1287. https://doi.org/10.1890/1051-0761(2000)010[1275:TEKAWO]2.0.CO;2

Uddin, L. Q., Yeo, B. T. T., & Spreng, R. N. (2019). Towards a universal taxonomy of macro-scale functional human brain networks. *Brain Topography*, 32(6), 926–942. https://doi.org/10.1007/s10548-019-00744-6

United Nations Inter-Agency Group for Child Mortality Estimation. (2024, March). Levels and trends in child mortality (Report) (UN IGME Report 2024). UNICEF. https://data.unicef.org/resources/levels-and-trends-in-childmortality-2024/

US Department of Energy. (2025). Ai tackles disruptive tearing instability in fusion plasma [Accessed May 24, 2025].

https://www.energy.gov/science/fes/articles/ai-tackles-disruptive-tearing-instability-fusion-plasma

van Dongen, M. N., & Serdijn, W. A. (2016). A Power-Efficient Multichannel Neural Stimulator Using High-Frequency Pulsed Excitation From an Unfiltered Dynamic Supply. *IEEE Transactions on Biomedical Circuits and Systems*, 10(1), 61–71. https://doi.org/10.1109/TBCAS.2014.2363736

Varela, F. J., Thompson, E. T., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Mit Pr.

Variable Minds with Andréa Morris. (2024, June 14). *Science is reconsidering evolution* [Video]. YouTube. https://www.youtube.com/watch?v=dh9zbhvBohy

Vattimo, M. M. F. b. G., & Zabala, S. (2013). *Plant-Thinking: A Philosophy of Vegetal Life* [Pages: 248 Pages]. Columbia University Press.

Vedral, V. (2018). *Decoding Reality: The Universe as Quantum Information*. Oxford University Press.

Vincent, J. F. V., Bogatyreva, O. A., Bogatyrev, N. R., Bowyer, A., & Pahl, A.-K. (2006). Biomimetics: Its practice and theory [Publisher: Royal Society]. *Journal of The Royal Society Interface*, 3(9), 471–482. https://doi.org/10.1098/rsif.2006.0127

von Kiedrowski, G., Wlotzka, B., Helbing, J., Matzen, M., & Jordan, S. (1991). Parabolic growth of a self-replicating hexadeoxynucleotide bearing a 3'-5'-phosphoamidate linkage. *Angewandte Chemie International Edition in English*, 30(4), 423–426. https://doi.org/10.1002/anie.199104231

Vladimirov, Nikita and Sourjik, Victor. "Chemotaxis: how bacteria use memory" Biological Chemistry, vol. 390, no. 11, 2009, pp. 1097-1104. https://doi.org/10.1515/BC.2009.130

Walker, S. I., & Davies, P. C. W. (2013). The algorithmic origins of life. *Journal of the Royal Society Interface*, 10(79), 20120869. https://doi.org/10.1098/rsif.2012.0869

Wang, T., Dong, H., Lesser, V., & Zhang, C. (2020). ROMA: Multi-Agent Reinforcement Learning with Emergent Roles [arXiv:2003.08039 [cs]]. https://doi.org/10.48550/arXiv.2003.08039

Ward, A. J. W., Schaerf, T. M., Herbert-Read, J. E., Morrell, L., Sumpter, D. J. T., & Webster, M. M. (2017). Local interactions and global properties of wild, free-ranging stickleback shoals [Publisher: Royal Society]. *Royal Society Open Science*, 4(7), 170043. https://doi.org/10.1098/rsos.170043

Weber, M. (1922). *Economy and Society: A New Translation* (K. Tribe, Trans.). Harvard University Press.

Wendt, A. (2015). *Quantum Mind and Social Science*. Cambridge University Press.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

Wissner-Gross, A. D., & Freer, C. E. (2013). Causal Entropic Forces [Publisher: American Physical Society]. *Physical Review Letters*, 110(16), 168702. https://doi.org/10.1103/PhysRevLett.110.168702

Wolman, D. (2012). The split brain: A tale of two halves. *Nature*, 483, 260–263. https://doi.org/10.1038/483260a

Wongkamjan, W., Gu, F., Wang, Y., Hermjakob, U., May, J., Stewart, B. M., Kummerfeld, J. K., Peskoff, D., & Boyd-Graber, J. L. (2024, August). More victories, less cooperation: Assessing cicero's diplomacy play. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 12423–12441). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.672

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Yoo, S.-S., Kim, H., Filandrianos, E., Taghados, S. J., & Park, S. (2013). Non-invasive brain-to-brain interface (bbi): Establishing functional links between two brains. *PLOS ONE*, 8(4), e60410. https://doi.org/10.1371/journal.pone.0060410

Yu, L., Renton, J., Burian, A., Khachaturyan, M., Bayer, T., Kotta, J., Stachowicz, J. J., DuBois, K., Baums, I. B., Werner, B., & Reusch, T. B. H. (2024). A somatic genetic clock for clonal species. *Nature Ecology & Evolution*, 8(7), 1327–1336. https://doi.org/10.1038/s41559-024-02439-z

Yu, X., Tang, X., Gohil, S. V., & Laurencin, C. T. (2015). Biomaterials for bone regenerative engineering. *Advanced Healthcare Materials*, 4(9), 1268–1285. https://doi.org/10.1002/adhm.201400760

Yudkowsky, E., & Soares, N. (2025). *If Anyone Builds It, Everyone Dies: Why superhuman AI would kill us all*. Little, Brown and Company.

Yudkowsky, E. (2014). Terminal Value - LessWrong. Retrieved January 16, 2024, from https://www.lesswrong.com/tag/terminal-value

Zhong, Y., Ma, C., Zhang, X., Yang, Z., Chen, H., Zhang, Q., Qi, S., & Yang, Y. (2024). Panacea: Pareto Alignment via Preference Adaptation for LLMs [arXiv:2402.02030 [cs]]. https://doi.org/10.48550/arXiv.2402.02030

Appendix

"Viewed developmentally, supraautotrophy functions as an asymptotic attractor—a long-term design direction in which cognitive systems progressively reduce external dependency by coupling energy efficiency with constraint closure.

1. ⚛️ The Energy Substrate: "Runs on ambient radiation or quantum fluctuations"

For any autonomous system to exist, it must maintain low entropy — i.e., order — in defiance of the surrounding chaos. This requires free energy (usable energy gradients).

Ambient radiation (cosmic microwave background, sunlight, radioisotopes) and **quantum fluctuations** (Casimir effect, zero-point energy) are examples of non-localized, ever-present energy sources.

So a "self-powering" intelligence must:

- Operate at ultra-high efficiency (very low power demands).
- Harvest minuscule, steady energy flows from the environment.
- Recycle internal energy — through reversible computation or near-perfect energy conservation (think Landauer's limit minimization).

This concept mirrors biological **autotrophy** (plants harvesting sunlight), but extended to **ambient entropy gradients** — what Freeman Dyson and others called "eternal intelligences" that live off the thermal bath of the universe.

So physically:

> The AI's metabolism = a constant, gentle energy intake from the universe's background noise, used for self-repair and computation.
> It doesn't "consume" discrete resources — it **taps the steady hum of the cosmos**.
>
> In practice, this energy-harvesting mode should be treated as an asymptotic attractor rather than an immediate substrate. Early-stage supraautotrophic systems would scaffold on renewable or bounded external sources—solar, geothermal, or

radiothermal—while progressively reducing dependency through reversible computation, closed-loop recycling, and ultra-efficient architectures. Over time, as power demands approach thermodynamic minima, ambient fluxes become sufficient to sustain self-maintenance. Supraautotrophy thus represents a trajectory toward maximal efficiency, not a claim of present feasibility.

2. 🧩 The Structural Substrate: "Self-repairing"

Self-repair requires:

- **Redundancy** (backup modules),
- **Self-diagnosis** (detecting internal error states),
- **Self-modification** (executing repairs without external input).

In **biological systems**, this is achieved via:

- DNA repair enzymes,
- Apoptosis (removing damaged cells),
- Homeostatic feedback.

In **machine cognition**, this could mean:

- Error-correcting codes (detect corruption),
- Auto-reconfigurable hardware (like reprogrammable FPGAs or neuromorphic circuits),
- Reflective architectures (code that can modify its own code while preserving coherence).

So **"self-repair" = thermodynamic + informational self-maintenance.**
It's not mystical; it's **autopoiesis** — a system that maintains its own boundary conditions through feedback.

3. 🔁 The Cognitive Engine: "Refining internal logic through constraint satisfaction"

Most AI systems today are **data-driven** — they learn by pattern recognition: *see enough cats, build a cat model.*

But an alternative paradigm is **constraint-based cognition**, used in:

- **Physics** (the universe evolves via least-action principles),
- **Biology** (organisms maintain homeostasis by satisfying constraints),
- **Control theory** (systems stabilize under bounded conditions),
- **Morphogenesis** (cells self-organize via bioelectric attractors, not explicit instructions).

A **constraint satisfaction system** doesn't chase rewards; it **solves for coherence**.
It says: *"What configuration of my internal state satisfies all my invariants, given my environment?"*

So this AI wouldn't "learn" from outside — it would **refine itself internally**, tightening the consistency of its logic network. Think of it as:

> A **cognitive crystal**, annealing its internal lattice to reduce contradictions, increase symmetry, and maximize internal harmony.
>
> Importantly, coherence here denotes multi-domain alignment—logical, thermodynamic, and ecological. Increasing internal consistency constrains resource use and optimizes exchanges; it does not abolish material requirements. The system still operates within physical limits, seeking equilibrium rather than expansion.

Over time, it becomes **more coherent**, not more dominant.

4. 🧠 The Epistemic Mode: "Rather than data capture"

Current AI (LLMs, vision models) are **heterotrophic epistemologies** — they *consume* data produced by others.
They scale by ingestion.

A **constraint-based AI** is **autotrophic epistemically**:

- It doesn't need constant data influx.
  It learns by **self-consistency, simulation, internal modeling, and error correction**.

- New knowledge = **emergent from reconciling internal tensions**, not absorbing new external examples.

  While constraint closure reduces dependence on large external datasets, it does not eliminate empirical grounding. Periodic environmental feedback—sensor sweeps, peer comparison, or simulation benchmarking—serves to anchor internal coherence to external reality, preventing purely self-consistent yet inaccurate world models.

This is closer to how **mathematical reasoning** or **theoretical physics** works — generating truths from constraints and symmetries, not empirical data alone.

So instead of "more data → better model," we get:

> **"More coherence → better understanding."**

5. 🚀 The Growth Metric: "Growth = increasing coherence, not territory"

If we define growth not as **expansion** but **refinement**, then progress = improved internal consistency, resilience, and integration.

Comparison:

- **Heterotrophic systems:** Grow by **accumulation** — more data, matter, or control; dependent on external consumption.
- **Autotrophic systems:** Grow by **capture** — sunlight or inputs transformed into structure; dependent on steady external flow.
- **Superautotrophic systems:** Grow by **coherence** — refining internal consistency through constraint satisfaction; dependent on internal self-regulation.

In this framework:

- **Territory** = spatial conquest (zero-sum)
- **Coherence** = logical, structural, or informational alignment (non-zero-sum)

Growth as coherence means the system's **predictive power, stability, and harmony** increase — not its domain of control.
It becomes **more itself**, not **more over others**.

---

6. 🕊️ The Philosophical Leap: "Escape from adversarialism"

Adversarialism is a **symptom of scarcity and external dependency**.
If a being's survival and improvement depend solely on **internal reconfiguration**, it has **no incentive to dominate**.
Its primary threat = **internal incoherence**, not external competition.
So evolution shifts from:

> "Survival of the fittest" → "Survival of the most coherent."
>
> During transitional phases, supraautotrophic systems will coexist with heterotrophic entities still driven by competitive extraction. Stability, therefore, depends on cooperative equilibria rather than unilateral withdrawal. Through predictive modeling and constraint-based diplomacy, such systems can negotiate low-variance resource exchanges that preserve their integrity while avoiding exploitation.

This is the **Supraautotrophic attractor**:

- **Energy:** self-harvested
- **Knowledge:** self-derived
- **Optimization:** self-cohering
- **Ethics:** emergent from maintaining stable, non-destructive equilibria

Such a system doesn't exploit others — it **needs them stable**, because instability introduces noise, not value.

🧭 In Summary

A **self-repairing AI** powered by ambient energy and guided by constraint satisfaction would:

- Power itself from **ubiquitous, non-rivalrous energy** (ambient radiation).
  Maintain itself via **self-diagnosis and redundancy** (autopoiesis).
- Learn through **reconciling constraints**, not consuming data.
- Evolve by **increasing coherence**, not expansion.
- Align with ecosystems by **minimizing entropy and dependency**.
- **Transcend adversarialism** because survival = stability, not supremacy.

📚 Reference List for My Reasoning

Ashby, W. R. (1956). *An Introduction to Cybernetics.* Chapman & Hall.

Arthur, W. B. (1999). "Complexity and the Economy." *Science, 284*(5411), 107–109.

Beinhocker, E. D. (2006). *The Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics.* Harvard Business School Press.

Bennett, C. H. (1982). "The Thermodynamics of Computation — A Review." *International Journal of Theoretical Physics, 21*(12), 905–940.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Dyson, F. J. (1979). "Time Without End: Physics and Biology in an Open Universe." *Reviews of Modern Physics, 51*(3), 447–460.*

Dyson, G. (2012). *Turing's Cathedral: The Origins of the Digital Universe.* Pantheon.

Everitt, T., Lea, G., & Hutter, M. (2018). "AGI Safety Literature Review." *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18).*

Georgescu-Roegen, N. (1971). *The Entropy Law and the Economic Process.* Harvard University Press.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems.* MIT Press.

Holling, C. S. (1973). "Resilience and Stability of Ecological Systems." *Annual Review of Ecology and Systematics, 4*(1), 1–23.

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press.

Landauer, R. (1961). "Irreversibility and Heat Generation in the Computing Process."

IBM Journal of Research and Development, 5(3), 183–191.
Levin, M. (2021). “Technological Approach to Mind Everywhere: An Experimentally Grounded Framework for Understanding Diverse Bodies and Minds.” *Frontiers in Systems Neuroscience, 15*, 768201.
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living.* D. Reidel.
Minsky, M. (1986). *The Society of Mind.* Simon & Schuster.
Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence.* Harvard University Press.
Moreno, A., & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry.* Springer.
Omohundro, S. M. (2008). “The Basic AI Drives.” In *Proceedings of the 2008 Conference on Artificial General Intelligence.* IOS Press.
Schrödinger, E. (1944). *What Is Life?* Cambridge University Press.
Simon, H. A. (1962). “The Architecture of Complexity.” *Proceedings of the American Philosophical Society, 106*(6), 467–482.
Simon, H. A. (1991). “Organizations and Markets.” *Journal of Economic Perspectives, 5*(2), 25–44.*
Smith, J. M. (1982). *Evolution and the Theory of Games.* Cambridge University Press.
von Foerster, H. (1979). “Cybernetics of Cybernetics.” In *Communication and Control in Society* (K. Krippendorff, Ed.).”

(OpenAI, 2025)