

Cognitive Infrastructures

Conjectural Explorations of AI as a Physical Actor in the Wild

Abstract

In this paper, we explore several conjectural scenarios for the development of artificial intelligence as a force animating the physical infrastructures of everyday life. We do so by foregrounding the real and potential coevolution of natural and machine intelligence in relation to one another. The proposals are diverse and developed semi-independently in the context of a one-month intensive research studio. If they are imagined as a single integrated speculation described from different vantage points, this speculation might be formulated as follows: the executable embedding of a society of digital twins of brain organoids that run in slow computational time by driving artificial epidermal mediated distributed organs, in turn part of xenomorphic biohybrid robotic phenotypes that produce various levels of nested minimum viable interiorities with which human users deploy inverted embedding visualizations to learn new game-theoretic dynamics through adaptive anti-transitivity, shifting asymmetries of mutual mind modeling, and carefully calibrated novelty-inducing disalignments. The composite assemblage would be fixed in a long-term durational imprint backed up on a highly durable substrate for long-term decoding.

Keywords

synthetic intelligence; artificial intelligence; cognitive infrastructures; productive disalignment; robotics; philosophy of artificial intelligence

**Cognitive Infrastructures:
Conjectural Explorations of AI as a Physical Actor in the Wild**

As AI becomes both more general and more foundational, it shouldn't be seen as a disembodied virtual brain. It is a real, material force. AI is increasingly embedded into the active, decision-making systems of real-world systems. As AI becomes infrastructural, infrastructures become intelligent, and as societal infrastructures concurrently become more cognitive, the relation between AI theory and practice needs realignment.

Natural Intelligence emerges at an environmental scale and in the interactions of multiple agents. It is located not only in brains but in active landscapes. Similarly, artificial intelligence is not contained within single artificial minds but extends throughout the networks of planetary computation: It is baked into industrial processes, it generates images and text, it coordinates circulation in cities, and it senses, models, and acts in the wild.

This represents an infrastructuralization of AI, but also a "making cognitive" of both new and legacy infrastructures. These new systems are capable of responding to us, to the world, and to each other in ways we recognize as embedded and networked cognition. AI is physicalized, from user interfaces on the surface of handheld devices to deep below the built environment. As we interact with the world, we retrain model weights, making actions newly reflexive in knowing that performing an action is also a way of representing it within a model. To play with the model is to remake the model, increasingly in real time.

What kind of design space is this? What does it afford, enable, produce, and delimit? When AIs are simultaneously platforms, applications, and users, what are the interfaces between society and its intelligent simulations? How can we understand AI Alignment not just as AI bending to society but also as how societies evolve in relationship to AI? What kinds of Cognitive Infrastructures might be revealed and composed? Across scales—from world-datafiction and data visualization to users and UI, and back again—many of the most interesting problems in AI design are still embryonic.

How might this frame human-AI interaction design? What happens when the production and curation of data is for increasingly generalized, multimodal, and foundational models? How might the collective intelligence of generative AI make the world not only queryable, but recomposable in new ways? How will simulations collapse the distances between the virtual and the real? How will human societies align toward the insights and affordances of artificial intelligence, rather than AI bending to human constructs? Ultimately, how will the inclusion of a fuller range of planetary information, beyond traces of individual human users, expand what counts as intelligence?

Individual users will not only interact with big models, but multiple combinations of models will interact with groups of people in overlapping combinations. Perhaps the most critical and unfamiliar interactions will unfold between different AIs, without human interference. Nascent ecologies are forming, framing, and evolving a new ecology of planetary intelligence.

The research is divided into five thematic sections:

1. Productive Disalignments
2. Post-Anthropocene Psycho-Physiologies
3. Organs Without Bodies
4. Planetary Time Computation
5. Mimesis of Mimesis.

Benjamin Bratton
Antikythera Director



1 Productive Disalignments

Complex intelligence arises from interactions among diverse minds, each shaped by unique priors, thinking styles, and communication modalities. Thus, the long-term evolutionary trajectory of artificial intelligence (AI) cannot be guided solely by the objective of alignment, particularly if alignment entails training AI to mirror human cognition closely. Instead, AI's potential for genuine innovation hinges precisely on its capacity to think orthogonally—to diverge meaningfully from human cognitive frameworks. This capacity positions AI as an “existential technology,” in the sense articulated by Stanisław Lem: a technology fundamentally capable of redefining our conceptual boundaries.

Reflectionism—the assumption that AI must reflect human cognition or be engineered strictly according to human-like parameters—has repeatedly driven discourse into conceptual and practical impasses. In contrast, *productive disalignment* emphasizes the value inherent in uncertain calibrations of novelty, alienation, and the unexpected pathways of coevolution between natural and artificial intelligences.

The notion of *productive disalignment* underscores the importance of allowing AI to develop and interact through cognitive paradigms that are intentionally distinct from human norms, creating dynamic potentials for innovation. The following papers delve deeper into this intricate balance by examining methods for measuring subjective novelty in generative AI outputs, alongside the processes of counteradaptation occurring between human and artificial minds. Together, these analyses illuminate the creative tensions essential for fostering meaningful and emergent forms of intelligence, highlighting productive disalignment as a critical guiding principle in the ongoing evolution of artificial cognition.

1a *Traversing the Uncanny Ridge*

Generative AI models, despite their vast creative potential, face a paradoxical challenge: the risk of “overalignment,” a phenomenon wherein generated outputs aesthetically collapse toward overly familiar norms, resulting in mundane, predictable images. This condition, which this paper terms the “Canny Valley,” is characterized by images that are eerily familiar—markedly different from Masahiro Mori’s “uncanny valley,” where discomfort arises from unfamiliarity. The Canny Valley represents a hyperconvergence between user expectations and generated outcomes, diminishing novelty and restricting creative exploration.

Addressing this issue, this paper introduces the concept of the “Uncanny Ridge,” an optimal zone of novelty and complexity where generative outputs evoke productive misrecognition, calibrated to stimulate curiosity and innovation without alienating the observer. Situated precisely between complete predictability and unrecognizable randomness, the Uncanny Ridge functions analogously to a “Goldilocks zone,” balancing maximum creative novelty against cognitive accessibility.

Recognizing that novelty is inherently subjective and context-dependent, influenced heavily by individual user priors, we propose a novel quantification framework in which novelty itself acts as a loss function. This mathematical formulation aims to operationalize novelty, enabling precise indexing and measurement tailored to varied user experiences and expectations.

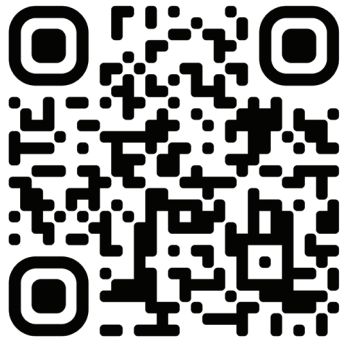
Further, drawing on insights from the psychology of creativity—specifically the capacity to hold contradictory ideas simultaneously—we suggest that sustained novelty emerges from dynamic tensions rather than simplistic divergence. Ultimately, the paper explores whether generative AI’s pursuit of novelty leads toward a productive Lagrange point of creative convergence or risks conceptual collapse. Navigating these intricate dynamics, it offers practical strategies for maintaining vibrant, meaningful innovation in generative AI.

1b *Synthetic Counteradaptation*

Artificial intelligence will not merely reflect human cognition; rather, it will profoundly reshape the trajectory of human thought itself, driving reciprocal adaptation between human and machine minds. This dynamic interplay of adaptation and counteradaptation raises critical questions about the mutual evolution of diverse cognitive systems. This project investigates this accelerated adaptive dialogue, particularly emphasizing convergence—the intriguing ways humans adapt to AI-learning processes even as AI simultaneously learns from human cognition. Fundamental to this inquiry is understanding how two distinct types of minds continuously recalibrate in response to one another. Biological evolution offers precedents such as predator–prey dynamics, where the adaptive strategies of one entity trigger counterstrategies in another, driving both toward escalating complexity. Another illustrative model is the strategic cycle inherent in games such as rock-paper-scissors, where success requires continuous predictive countermodeling of an opponent’s thought processes.

Adaptation strategies can broadly take two forms: mirroring or deceiving the opposing mind. Notably, the Turing Test exemplifies both, employing mirroring as a sophisticated form of deception aimed at convincing humans of machine authenticity. Likewise, AlphaGo’s landmark interactions—specifically its celebrated Move 37 and Lee Sedol’s Move 78—highlight adaptive anti-transitivity. Move 37 was a groundbreaking strategy by AlphaGo, initially seen as puzzling or incorrect by human experts but later revealed as ingeniously creative. Move 78, executed by Lee Sedol in response, similarly defied conventional wisdom and showcased an unprecedented human adaptation to the AI’s unconventional strategy. These moves illustrate how novel and initially perplexing decisions emerge through intricate layers of reciprocal mind modeling and anticipation.

Ultimately, by exploring these adaptive dynamics, this project maps how human and artificial intelligences will coevolve. Through such mutual cognitive reshaping, it proposes that the future will be defined not merely by AI’s capacity to imitate human thinking but equally by humanity’s profound and nuanced adaptation to the evolving logic and learning patterns of AI.



Traversing the Uncanny Ridge

Searching for Novelty in Intersystemic Communication

Sonia Bernac
Royal College of Art
London

Tyler Farghly
University of
Oxford

Gary Zhexi Zhang
Independent
Researcher

Abstract

This paper develops the uncanny ridge as a thought experiment within debates on AI alignment, intersystemic communication, and the limits of novelty in machine learning. Repositioning systemic misalignment in AI as a generative rather than disruptive force, it draws on and inverts Masahiro Mori's uncanny valley, which describes human discomfort toward near-human simulations. The uncanny ridge shifts focus from breakdowns in recognition to the conditions under which misrecognition between AI systems produces novelty. While alignment is often assumed to improve coherence, excessive synchronization leads to hyper-convergence, constraining AI's capacity for emergent intelligence. To formalize this, the paper introduces the uncanny index, a metric for identifying when systemic misrecognition produces innovation rather than collapse. Case studies of mode collapse in generative adversarial networks (GANs), reinforcement learning with human feedback (RLHF) constraints, and multimodal AI architectures demonstrate that novelty does not emerge from complexity alone but from structured divergence and delayed alignment. Rather than treating misalignment as a failure, this thought experiment reframes the uncanny ridge as a dynamic site of intersystemic tension, where AI models engage in productive misrecognition, generating new computational strategies and unexpected modes of coordination.

Keywords

artificial intelligence; uncanny valley; uncanny ridge; complexity; novelty; intersystemic communication; AI infrastructures

1 Introduction

The assumption that alignment—whether between humans and AI or between AI models themselves—is inherently beneficial not only obscures the investigation into the ontology of synthetic intelligence but also constrains the design of planetary-scale interactions between diverse AI systems, operating across different architectures, logics, and modes of processing.

The ongoing exploration of the differences between humans and AI often leads to a conflation of diverse algorithmic systems into a singular, undifferentiated synthetic “other.” However, AI models, platforms, and tools in the wild do not form a unified, synchronized synthetic layer, especially when woven into the pattern of legacy infrastructures. Instead, the development of *cognitive infrastructures*—the technical architectures of AI, data processing, and planetary computation embedded in physical environments—reveals ecologies of synthetic intelligences that operate with fundamentally different languages, technological protocols, timescales, and sensory inputs and outputs. The friction and tensions between these data bodies, models, and sensory domains can either be treated as issues to be resolved or embraced as opportunities to fundamentally rethink infrastructural design needs.

Crucially, such a diverse synthetic ecosystem is rarely reflected in the outputs associated with AI-generated content. Despite the variety of scenarios in which they are employed, these outputs frequently exhibit a derivative, recombinatorial quality—a form of canny familiarity.¹ Rather than concluding that AI is therefore a “stochastic parrot,” this paper identifies the systemic factors contributing to the convergence of meaning across different modalities, resulting in predictable outcomes.²

Convergence is neither essential nor always desirable for effective collaboration. Interactions within and between systems—whether biological, technological, or social—often flourish through the tensions and conflicts that arise from diverging perspectives. Robustness, adaptability, and flexibility in complex systems frequently depend on delayed alignment, the presence of noise, and varied interpretations of the system's functioning or goals.³ Moreover, misinterpretation, miscommunication, and mistranslation have historically driven the emergence of new patterns in evolution and civilization.⁴ Complete systemic transparency (informational alignment) can lead to informational overload, while full synchronization (operational alignment) may stifle the diversification of roles vital for efficiency. Thus, selective functional signaling, rather than full transparency, is often key to successful task completion. While alignment may seem advantageous, it can sometimes undermine the very conditions necessary for innovation.

1.1 Intersystemic Interactions in Machine Learning

The drive toward developing foundation models—systems capable of handling multiple modes of input and output—reflects a broader ambition within the machine learning community to create more versatile and generalized models.⁵ However, this ambition is currently constrained by two significant challenges. First, the architecture of a model is often tightly coupled to the structure of the data they are trained on, making it difficult to incorporate multiple modes of input. Second, these models typically require supervision with large datasets, and creating datasets that encompass all modes of input and output is inherently difficult.⁶ A proposed solution to this problem involves training individual machine learning models on different types of input separately and then integrating these models to create a more comprehensive system.

An example of this approach is contrastive language-image pretraining (CLIP), which seeks to unify visual information from images with linguistic information from image labels. It works by having language models and image models occupy the same embedding space and by having the embeddings of these two forms of data—image and text—converge on this space. By creating this shared space that bridges different modes of information, downstream image tasks can be performed with smaller datasets than previously possible.⁷

Diffusion models are a class of generative models that have rapidly gained popularity, partially due to their ability to combine information from pretrained models. For instance, an image classifier or a model like CLIP can be used to influence the output of a diffusion model according to a specific prompt or conditioning. This process, known as diffusion guidance, is performed by forcing the diffusion model to agree with the pretrained model on the feasibility of the output. This capability of prompt-based data generation is largely responsible for the widespread adoption of diffusion models and their integration with language models. The ability to leverage interaction between models trained on different tasks not only enhances the performance of these systems but also extends their applicability across a wider range of tasks.

¹ Bernac and Keenan, “Diffusion Models.”

² Bender et al., “Dangers of Stochastic Parrots.”

³ Bouffanais, *Swarm Dynamics*.

⁴ Schmutzer and Wagner, “Mistranslation Alters.”

⁵ Bommasani, “Opportunities and Risks.”

⁶ Bommasani et al., “Opportunities and Risks.”

⁷ Kim et al., “DiffusionCLIP.”

However, CLIP-guided diffusion models have been criticized for producing unremarkable outputs, displaying well-known aesthetics and commonsensical semantic associations while promoting dataset memorization.⁸ This method allows for generalization across a wide range of tasks, but it also limits the model's exploratory capabilities. By operating from a matrix of pre-clustered statistically common associations found in its training data, CLIP guidance leads to content homogenization and alignment with conventional rather than novel styles.

Within intersystemic interaction, human feedback serves as a form of intelligent agent input, shaping the alignment and operational dynamics of AI systems. RLHF is a training method used to fine-tune generative models by incorporating human evaluations into the learning process. It works by creating an auxiliary reward model that helps an AI-in-training predict user satisfaction. This reward model is then used to optimize the system, guiding it to produce outputs that maximize alignment with human preferences.

RLHF has undoubtedly been instrumental in shaping language models to meet the demands of economic utility, yet this very success has come at the cost of creative diversity. The phenomenon of mode collapse, where the richness of possible outputs is severely diminished, is a direct consequence of this alignment process.⁹ Research has shown that systems closely guided by human feedback tend to replicate existing styles and conventions rather than exploring new creative directions.¹⁰ This phenomenon underscores the trade-off between aligning with human expectations and maintaining a system's capacity for innovation. This outcome, however, should not be surprising; RLHF optimizes for conformity with simplified models of average human satisfaction—an approach that is almost inherently opposed to the vibrancy and diversity that characterizes genuine novelty. RLHF also introduces significant constraints on the generative potential of machine learning systems by reinforcing existing biases and human-centric expectations.¹¹ This constraint is particularly evident in studies of RLHF applied to content moderation and recommendation systems, where the goal is to minimize user discomfort and maximize engagement.¹² Human feedback tends to reward familiarity and penalize deviation from expected norms. Over time, this feedback loop can lead to *preference drift*, where the diversity of outputs diminishes as the system becomes increasingly tuned to produce what users are most likely to accept.¹³ This effect has been observed in personalized content delivery systems, where algorithms prioritize content that aligns with established user preferences, thus limiting exposure to novel or challenging ideas.¹⁴ The resulting cognitive monoculture constrains the space of possible outputs, reducing the likelihood that the system will generate the unexpected or the strange.

The ideological foundations of AI alignment research seem to be at odds with, or entirely dismissive of, the potential for genuinely novel artificial intelligence. The underlying assumption is that AI should mirror human expectations, behaviors, and desires, leaving little room for the unpredictable or the unprecedented. In this framework, anything that deviates from the norm is seen not as a possible innovation but as an anomaly to be corrected. This ethos of alignment, then, risks stifling the potential for artificial intelligence to produce something truly new—something that does not simply reflect human tendencies but transcends them.

More broadly, consider the rapidly emerging ecologies of interaction that span from infrastructural systems—smart grids, autonomous logistics networks, algorithmic governance frameworks—to the more familiar realms of commercial and personal AIs, such as digital assistants, recommendation systems, and predictive algorithms. These agents, each designed with specific operational logics and purposes, are increasingly required to communicate not just with humans but with one another. The dialogues between autonomous vehicles and traffic management systems, the negotiations between financial algorithms in high-frequency trading, or the interactions between personalized AI agents in consumer markets, all exemplify a new space of system-to-system communication. This increasing complexity of AI cognitive infrastructures implies the emergence of unprecedented forms of interactions, distributions, and circulations that are not simply the sum of technological components.¹⁵

1.2 Complexity versus Novelty

This assumption that greater complexity automatically leads to greater novelty often arises from conflating distinct definitions of complexity, including the mistaken notion that complexity simply means greater complication. These varying meanings are then collapsed into an intuitive but misleading understanding of the term. Crucially, complexity is defined in distinct ways across scientific domains, including computational, algorithmic, information-theoretic, and systems-based frameworks. Computational complexity examines the resources, such as time and space, required to solve classes of

⁸ Wang et al., “Diffusion Feedback”; Somepalli, et al., “Copying in Diffusion Models.”

⁹ Kirk et al., “Effects of RLHF.”

¹⁰ Riedl and Harrison, “Using Stories.”

¹¹ Ho and Ermon, “Generative Adversarial Imitation Learning.”

¹² Brown et al., “Value Alignment Verification.”

¹³ Kingma and Welling, “Auto-Encoding Variational Bayes.”

¹⁴ Gillespie, “Relevance of Algorithms.”

¹⁵ ACS Research, “Alignment of Complex Systems.”

problems.¹⁶ Algorithmic complexity, based on Kolmogorov theory, measures an object's complexity by the length of the shortest algorithm capable of producing it.¹⁷ In information theory, complexity is often equated with entropy, representing the degree of uncertainty or randomness within a system.¹⁸ Complexity science, in turn, studies emergent behaviors arising from interactions among components, resulting in properties that cannot be reduced to or deduced from the sum of individual elements.¹⁹ While each of these frameworks captures different aspects of complexity, none offer a universal mechanism for generating novelty. Rather than prescribing a singular model, they demonstrate how novelty may arise under specific conditions, often in ways that resist direct prediction or deterministic formulation.

The definition of novelty is far from straightforward, as it shifts depending on context, criteria, and disciplinary perspective. Novelty remains a persistent philosophical problem, raising questions about the conditions under which something can be considered genuinely new, whether it is always contingent on prior structures and how its recognition is shaped by existing conceptual frameworks. In scientific and mathematical contexts, it is frequently understood through statistical deviation, unpredictability, or emergent patterns. However, this comes with the risk of conflating novelty with mere anomaly or randomness. In artistic and cultural theory, novelty is tied to aesthetic shifts or ruptures in the logics of making, yet what counts as novel is always contingent on existing frameworks of interpretation. These competing understandings make it difficult to formalize novelty as a stable category, a challenge that becomes even more pronounced when applied to AI-generated outputs.

Suppose we take complexity to be the number of components in a system and their interactions and contrast it with novelty, which we take to be the emergence of previously non-existent elements or configurations.²⁰ Then, as complexity increases, the potential for novelty seems to grow, offering more possibilities for new patterns to emerge. However, this relationship is intricate. Increasing the number of elements and interactions does not automatically lead to greater novelty, and a logical systemic leap does not always arise from complexity or exhibit complexity itself.²¹ Novelty can emerge from limited simple interactions or through a system's unique ability to reduce complexity (compression), challenging the assumption that more complexity inherently fosters more novelty. From this, it can be seen that there is no simple recipe for systemic novelty.

Discourses on novelty in AI often fall into essentialist debates over whether AI can produce anything truly new, or, conversely, succumbing to relativistic views that dismiss the possibility of true novelty altogether.²² Crucially, discourses surrounding AI-generated content often overlook insights from art and media theory. A broader modernist move, particularly evident in the mid-twentieth century, emphasized *processes* as integral to an artwork—moving beyond focus on the final artifact.²³ This shift highlights a perspective from which AI approaches could benefit, reorienting focus from output production to designing experimental architectures that challenge the creative process itself.²⁴ For example, if an AI model produces a predictable, uninspired outcome—such as a dull image—but does so through an unfamiliar or unconventional logic that reveals a new mode of reasoning or creation, this challenges conventional notions of novelty. Unfortunately, discussions about AI often reduce its dynamic processes to mere artifact production and equate novelty with the arbitrary perceptual surprise at its effects produced in human agents. In reality, novelty is ontologically perspectival, not because it requires an observer for validation, but because it exists within a bounded system with a particular history and memory that provide a framework for recognizing novelty. Novelty requires a systemic rather than individual perspective.²⁵

1.3 Recognition and Dissonance

The uncanny valley was originally conceptualized by Masahiro Mori in the context of human–robot interaction. The uncanny valley represents an epistemic rupture—a point at which anticipations of continuity confront a profound dissonance between expectation and reality. As machines approximate the human form, the comforting familiarity we experience is transformed into unease or outright horror when the semblance of life becomes too convincing yet remains distinctly other. This break is not just a glitch in perception; it is a moment of cognitive dissonance that reveals the precarious balance between recognition and misrecognition, between the familiar and the alien, thus exposing the fault lines in our understanding of what it means to be human.²⁶ This rupture is an emergent property of our cognitive architecture, a testament to the brain's propensity to resist that which challenges its categorizations.²⁷

¹⁶ Arora and Barak, *Computational Complexity*.

¹⁷ Li and Vitányi, *Introduction to Kolmogorov Complexity*.

¹⁸ MacKay, *Information Theory*.

¹⁹ Ladyman et al., "What Is a Complex System?"

²⁰ Standish, "Concept and Definition."

²¹ Felin and Kauffman, "Search Function."

²² Cf. Kraatz and Xie, "AI Art Is Not Art"; Chiang, "A.I. Isn't Going."

²³ Kaprow, *Art and Life*.

²⁴ Bishop, *Installation Art*.

²⁵ This is an original formulation inspired by the distinction between spatial complexity and temporal novelty found in Stang, "Novelty and Complexity."

²⁶ Mori, "Uncanny Valley."

²⁷ Saygin et al., "Predictive Coding."

The uncanny valley is traditionally framed as a phenomenon of human perception, but at its core, it reflects a more fundamental problem of recognition—one not limited to human observers but present in interactions between systems, models, and intelligences. One prevailing hypothesis is that the uncanny valley arises from a conflict between competing neural processes: a familiar but slightly distorted human-like figure may simultaneously activate cognitive pathways associated with both empathy and revulsion. The brain's mirror neuron system, which plays a role in recognizing and empathizing with others, might be disrupted when faced with an entity that appears human but exhibits non-human features or behaviors. This mismatch can trigger a threat response, rooted in evolutionary mechanisms that prioritize the detection of potential dangers posed by diseased or deceased conspecifics.²⁸

Further evidence from neurocognitive research suggests that the uncanny valley may be linked to the brain's predictive coding framework, where the brain continuously generates and updates mental models of the world. When an encountered entity deviates from expected human norms in subtle but significant ways, the prediction error generated may be too large to reconcile smoothly, leading to a sense of unease or eeriness.²⁹ This is particularly evident in cases where the entity's appearance or behavior is close enough to human to engage higher-level social cognitive processes but different enough to disrupt them, leading to an uncanny experience. This also explains why self-likeness plays a critical role in the phenomenon: the closer something is to being an accurate mirror of the self, the more sensitive the brain is to any discrepancies. This is amplified in our perceptual and emotional responses.

However, recognition is not merely a perceptual or cognitive issue—it is also a systemic one. Across intersystemic interactions, recognition is always partial, negotiated, and contingent on the dynamics of alignment and misalignment. Misrecognition, like perceptual dissonance, is often treated as a problem to be resolved, yet it is equally a driver of systemic novelty. Without a framework for recognition, the emergence of the truly new—an anomaly, a pattern break, a deviation, a logical leap—would be impossible. It is in these moments of dissonance, where familiar logics become unstable, that systems recalibrate, generate new pathways, or transform the conditions of intelligibility.

1.4 The Uncanny Ridge

Early iterations on GANs and diffusion models generated images of humans with grotesquely distorted features, like limbs twisted into unnatural shapes and fingers multiplying beyond recognition. In their nascent forms, generative AI models unintentionally produce novel imagery with strikingly unique aesthetic qualities. However, these early sparks of novelty were swiftly extinguished. The image models were rapidly updated to eliminate such artifacts, with these emergent, strange aesthetic qualities being dismissed as mere errors or deviations from some presumed ideal. Preserving novelty in these models requires resistance to the gravitational pull of optimization-driven redundancy, maintaining enough divergence to foster genuinely new behaviors. However, current architectures often fail to reach this space, as optimization pressures drive systems toward convergence, reinforcing existing patterns rather than allowing them to explore uncharted interactions. The challenge, then, is to structure machine learning models in a way that sustains the conditions necessary for systemic novelty.

In the conventional uncanny valley graph, the valley represents a trough of discomfort, where resemblance without full recognition produces unease. We propose to reconceptualize the topology of the uncanny valley, not as a dip in affective response but as a dynamic terrain where interactions between systems—biological, artificial, and hybrid—do not result in breakdowns of recognition but in the intersystemic emergence of novelty. The *uncanny ridge* marks a peak of interactional intensity—a zone where intentional miscommunication, misrepresentation, and misrecognition between intelligent agents can lead to perspectival novelty and significantly enhance outputs' originality. A point where the encounter between heterogeneous models and intelligences does not lead to failure or estrangement but instead generates new logics of engagement. At this peak, intelligibility is not merely stretched or distorted but actively reorganized, as AI and other non-human agents co-compose new forms of sense-making at the interstices of recognition and misrecognition, prediction, and unpredictability.

Within this framework of the uncanny, the potential for misalignment among intelligent systems takes on heightened significance. Misalignment is not merely a technical anomaly but a challenge to the very concept of a monoculture that can arise when systems become overly aligned or synchronized. This work proposes a formalization for identifying areas where productive differences can lead to systemic novelty, focused primarily on intersystemic interactions within machine learning architectures. By inspecting the nature of these interactions, as well as the introduction of the uncanny index, it attempts to examine novelty in systemic rather than anthropocentric terms, defining meaningful change through the lens of systemic prediction. This approach avoids the fetishization of misalignment as a universally desirable strategy, while simultaneously moving beyond human-centered interpretations of novelty, reframing it in terms of a system's own dynamics, predictive capacities, and internal logic.

²⁸ Eagleman, *Incognito*.

²⁹ Seth, "Cybernetic Bayesian Brain."

2 Mapping Intersystemic Interactions

In this section, we examine examples and forms of misalignment and map them through the lens of novelty. We begin by exploring distinct interaction protocols between pairs of AI systems (Figure 1). These protocols formalize how different patterns of communication and recognition can either constrain or expand the semantic space of these models. By systematically examining these misalignments, it becomes possible to observe the mechanisms behind semantic expansion, systemic change, and the emergence of novelty. Through this exploration, the section sheds light on how misalignment can lead to novel behaviors, unexpected innovations, or, conversely, the collapse of meaning within these systems. This framework lays the foundation for comprehending the broader implications of AI misalignment and its potential impact on cognitive infrastructures.

2.1 A Taxonomy of Misalignment

The simplest form of interaction between two intelligent systems is the *null protocol*, in which each system is probabilistically independent of the other. In this scenario, they exert no influence on one another, and there is no attempt at mutual modeling. This serves as a baseline for understanding the role of alignment on novelty, as the introduction of this form of interaction serves no bearing on the potential novelty of the pair of systems.

The second protocol that we introduce is the *intersection protocol*. In this scenario, the drive for the two systems to agree leads to a narrowing of the space of semantic possibilities. We refer to this kind of intersection within semantic space as *hyper-convergence*. An explicit example of this can be found in CLIP-guided diffusion models, where a battle between an unconditional diffusion model and CLIP leads to a literal convergence at the intersection of their supports. This also encompasses the tendency toward mode collapse brought about by RLHF, where agreement between language models and human satisfaction reward models can restrict diversity of output. Compared to the null protocol, this form of interaction evidently produces a collapse in novelty.

A third example of this can be found in the literature on GANs. This is a generative architecture formed of two models—generator and discriminator—that compete in a game in which the generator attempts to produce synthetic data that fools the discriminator.³⁰ A well-documented issue with these models is also a form of mode collapse, where the diversity of generated outputs contract, leading the system to converge on a limited subset of possibilities.³¹ This occurs when the generator in a GAN, whose task is to create data indistinguishable from real data, becomes overly optimized to produce outputs that consistently deceive the discriminator. This results in a reduction in the system’s ability to explore its full creative potential by narrowing the space of possible outputs. This collapse in novelty is not merely a technical flaw but reflects a deeper issue in how these systems are trained and optimized. As machine learning models increasingly prioritize accuracy and efficiency, they tend to overfit to the most successful patterns, thereby sacrificing the exploration of less probable but potentially more innovative outputs.³²

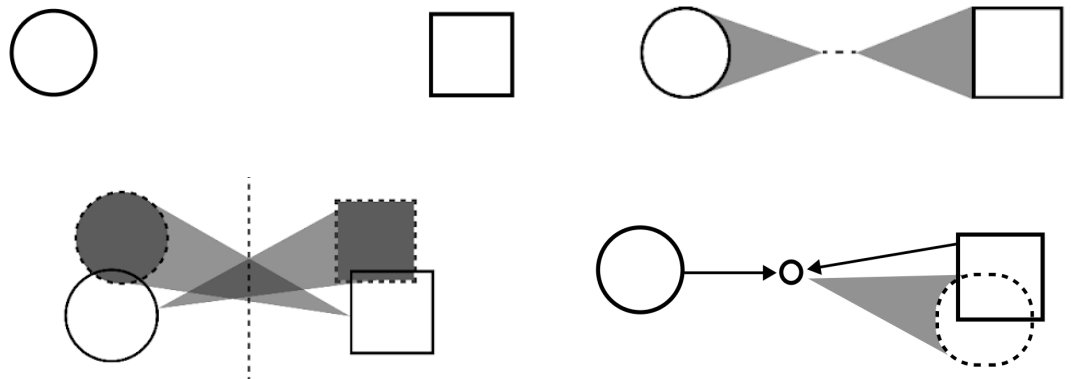


Figure 1 Diagrams of interaction protocols. Clockwise from the top-left: the null protocol, the intersection protocol, the mutual projection protocol, the Roadside Picnic protocol.

The third protocol we discuss implies a form of mutual misaligned projection, which we call the *mutual projection protocol*. In this context, two agents are capable of modeling one another or have a strong ability to predict each other’s activities, yet they cannot directly see each other or access each other’s realities. Despite this lack of direct visibility, they continue to form projections of one another.

³⁰ Goodfellow et al., “Generative Adversarial Nets.”

³¹ Lucic et al., “Are GANs Created Equal?”

³² Higgins et al., “beta-VAE.”

This can drive innovation and lead to strategies that diverge in novel directions, which would not have emerged if direct visibility were available.

An example of this can be found in Google’s DeepDream, which takes a deep neural network trained for an image classification task and forces it to generate images.³³ It does this by iterating on patterns detected in the model, enhancing them to the point where they become surreal or hallucinatory, producing outputs that are both familiar and bizarre. Studies of DeepDream have shown that this process can lead to the emergence of novel visual forms, which are strikingly different from the inputs on which the network was originally trained.³⁴ The strangeness produced by DeepDream is a form of generative novelty, where the familiar is rendered strange, inviting new interpretations and expanding the boundaries of visual creativity.

A more abstract example of this generative dynamic can be found in the Cold War, where a recursive loop of paranoia and projection materialized in the form of technological acceleration. Nuclear deterrence, cryptographic security, and aerospace expansion—each driven by the imperative to preempt an enemy—became engines of unintended innovation. Like DeepDream amplifying patterns into visual delirium, the Cold War exaggerated its own signals into material infrastructures and technological progress.

We can further lean into more abstract examples of interaction protocols by borrowing from fictional literature. For example, we consider the *Roadside Picnic protocol*, inspired by the Strugatsky brothers’ novel of the same name. In the novel, hazardous zones emerge after an unexplained extraterrestrial event, leaving behind enigmatic artifacts that, while seemingly trivial to their alien creators, have profound and unpredictable impacts on human civilization. In this protocol, one intelligent system generates an output, which another system—without direct access to the original context—attempts to interpret and integrate within its own language and operational framework. This indirect interaction enables the interpreting system to expand its semantic space, enhancing its interpretive range and adaptability.

2.2 Novelty and Misalignment

These exemplary protocols can be mapped on a simple four-quadrant diagram (Figure 2). On the top horizontal axis, the extremes are labeled “can’t see” and “can see,” while on the vertical axis, the labels are “can’t model” and “can model.” To “see” in this context refers to the ability to perceive phenomena related to the other system—to register change. On the other hand, to “model” implies having some form of understanding or predictive capability, similar to a “theory of mind,” or at the very least, a probabilistic mechanism capable of forecasting the other system’s next move.

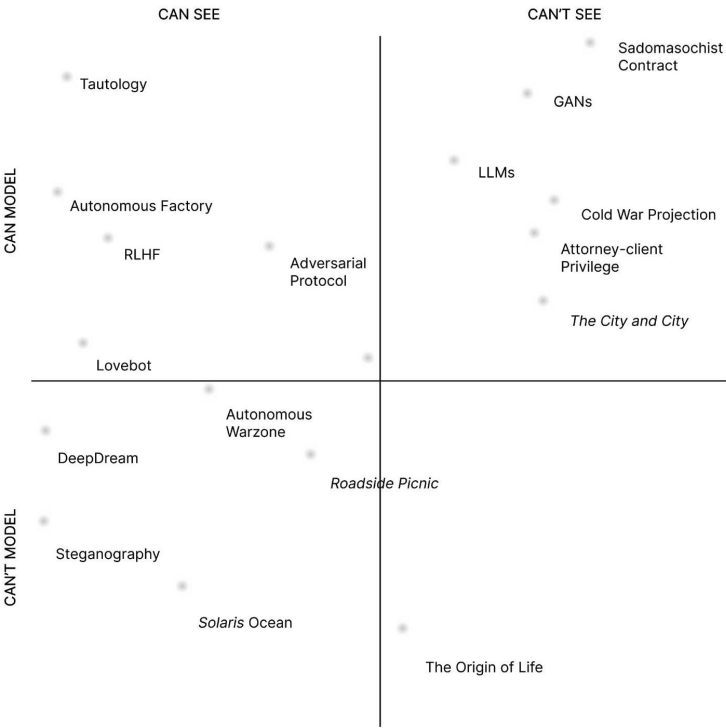


Figure 2 A two-dimensional mapping of interaction protocols.

³³ Mordvintsev et al., “Inceptionism.”
³⁴ Bronstein et al., “Geometric Deep Learning.”

To systematize these examples further, this paper introduces a vertical axis ranging from “perspectival novelty” at one extreme to “collapse of meaning” at the other. As novelty emerges along this spectrum, it resists straightforward prediction or proportional scaling, yet distinct dynamics become evident within each identified quadrant. Mapping these dynamics provides insights into the generative behaviors that result from interactions between different intelligent systems. Various examples of AI and non-AI systemic interactions are plotted on this diagram, showing where novelty increases or declines.

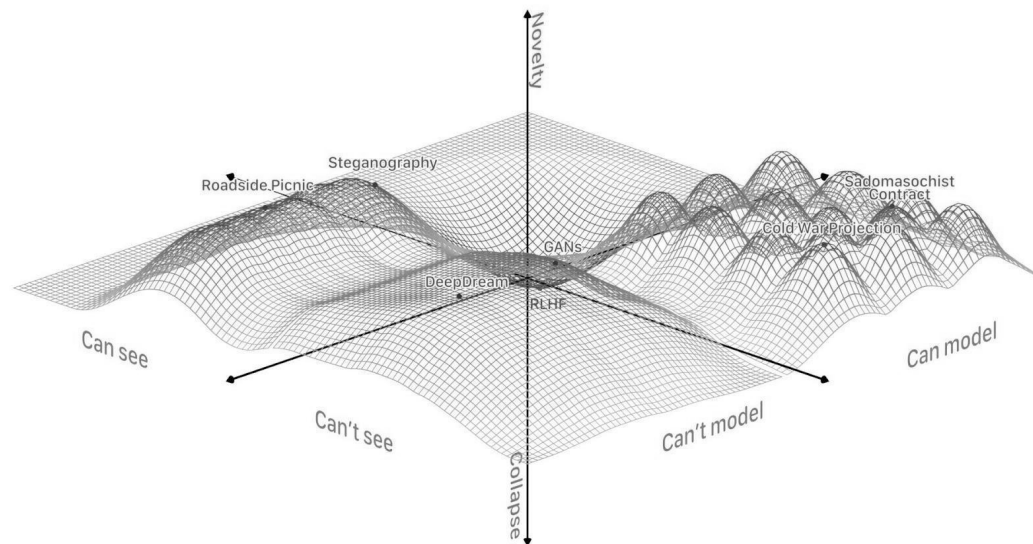


Figure 3 Plotting interaction protocols against their estimated effect on the promotion of novelty. An increase in the vertical axis represents an increase in systemic novelty.

For example, the can model–can see quadrant contains examples of the intersection protocol such as RLHF and GAN mode collapse, which, as we have noted, can lead to a convergence on the familiar, which we represent by a drop in the vertical axis. In the can’t see–can’t model quadrant, we have DeepDream, where interactions produce outcomes that challenge conventional visual norms, represented by an increase in the vertical axis (Figure 3).

3 Traversing the Uncanny Ridge

In this section, we define the uncanny ridge and speculate on how it might be traversed through the empirical measurement of novelty as well as the systematic exploration of taxonomies of misalignment.

3.1 The Uncanny Ridge

The uncanny ridge identifies the perspectival disruption of the uncanny valley across systems—biological, artificial, hybrid—as a topography of productive misapprehension and novelty. Where the valley marks discomfort, the *uncanny ridge* marks a peak of interactional intensity—a point where the encounter between systems generates new logics of engagement. At this peak, intelligibility is not merely stretched or distorted but actively reorganized through coevolution. Examples of this abound in nature, such as in plant–insect pollination mechanisms, as well as in human–computer interaction, where human response to AI models is often characterized by flawed heuristics for discerning computer-generated output.³⁵

Between models, this narrowing of generative diversity not only threatens model collapse but also restricts the potential for meaningful interaction between models, reinforcing alignment at the cost of emergent complexity. It is precisely within this tension—between optimization-driven collapse and the need for systemic divergence—that the uncanny ridge seeks to identify an optimal zone of maneuver. It is *topological* because it describes a structural dynamic between systems rather than a discrete transition. As AI proliferates, not as isolated artifacts but as infrastructural agents embedded in planetary-scale computation, the uncanny ridge provides a framework for understanding how selective misalignment enables systemic novelty—not as an anomaly but as an inherent feature of complex cognitive infrastructures.

In what follows, we sketch the experimental conditions for measuring this dynamic terrain empirically.

³⁵ Jakesch et al., “Human Heuristics.”

3.2 Proposed Experiment

We propose that novelty is perspectival, defined by its relation to specific systemic priors. Mapping and navigating the uncanny ridge requires an external or internal framework to identify the newness against those systemic priors. These priors are not always spatially or immediately present; they can also be temporally embedded as memories or tendencies developed through training.

To explore where novelty emerges within intersystemic interactions, we propose a thought experiment that identifies the conditions of misalignment under which the uncanny ridge can be mapped. To make these conditions explicit, we externalize the process by introducing an observer—either an external agent or an internal systemic component (represented separately for clarity in a diagram)—that tracks and assesses patterns of novelty in agents' outputs or behaviors across varying degrees of alignment. By plotting the detected levels of novelty against shifting alignment conditions, the uncanny ridge is mapped as a trajectory, highlighting where misalignment leads to emergent, unexpected outcomes rather than predictable or collapsed states (Figure 4). While this remains a theoretical model, it provides a structured framework for identifying where to look for novelty within differing misalignment conditions in future empirical studies.

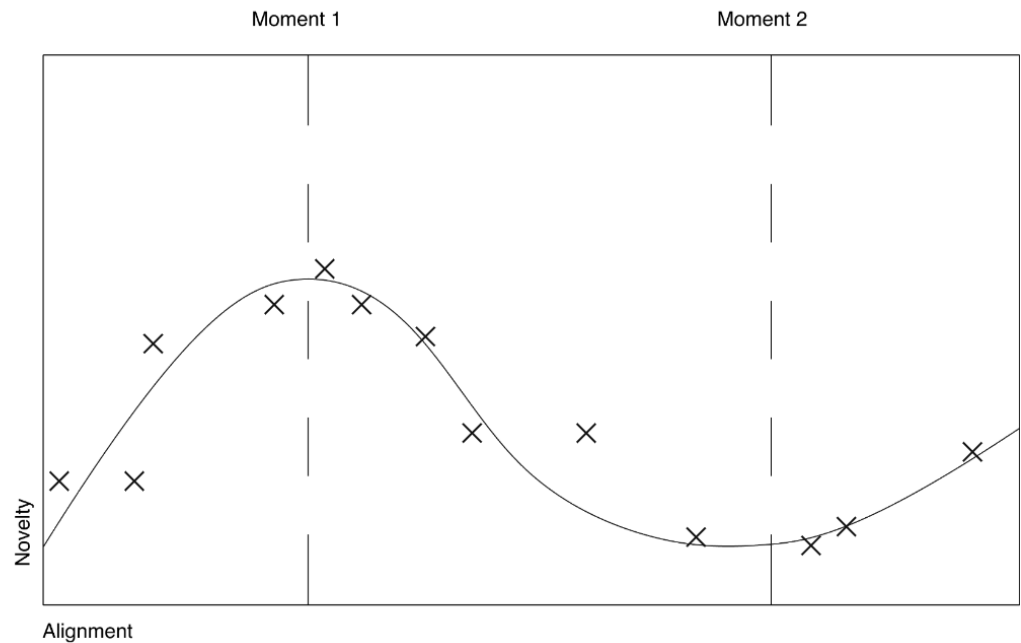


Figure 4 In the proposed experiment, we map novelty against different conditions of a given interaction protocol.

The measurement of novelty is influenced by several factors, including the depth of the agent's memory or the temporal a priori, the perceptual capabilities of the observing agent, the accessibility of the observed qualities, and the agent's positionality relative to the system (Figure 5). Thus, this experiment involves multiple measurements from different positions, observers, and frameworks of reference within the systems to identify potential overlaps and intersections. By aggregating these diverse perspectives, the goal is not to establish an average or universal novelty but to gain a deeper understanding of the dynamics between the two systems in the absence of a pre-established absolute alignment space, such as CLIP. The aim is to evaluate and dimension the space of misalignment and begin to formulate the various parameters that can exist within it.

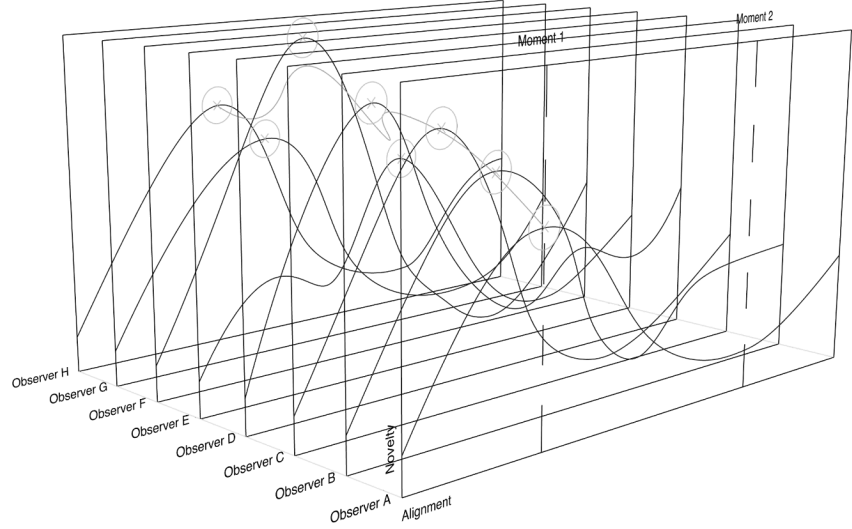


Figure 5 Novelty is perspectival. In our proposed experiment, we propose that novelty should be measured from the perspective of different observers and aggregated.

3.3 Measuring Novelty Increase

The empirical measurement of novelty occupies a critical role across various domains, particularly within cognitive science, where novelty is quantified by evaluating the deviation of a stimulus from an individual’s prior experiences or expectations. This quantification is often rooted in assessing how an individual assigns likelihood to a prediction, typically through its associated log-probability—a concept known as Bayesian surprise.³⁶ In transposing this notion of novelty to our present context, we measure the shift in surprise induced by the implementation of an interaction protocol, thereby constructing what we term the uncanny index. This uncanny index represents a metric for novelty that captures the expected change in log-probability between two scenarios: one in which a pair of AI systems interacts under a specific protocol, and another where this protocol is absent—the setting of the null protocol.

We can describe this in the language of probability theory, treating the output of a pair of systems as a random variable. Denote by \mathbf{p} the probability density of the prior distribution on some measurable space $X \times Y$. Then, given a pair of random variables X and Y on the spaces X and Y , the unnormalized surprise is given by the quantity,

$$\mathbb{E}[-\log p(X, Y)].$$

If X, Y is formed from the output of some pair of systems, then the surprise measures how much this output is to be expected under the prior distribution—a kind of perspectival and artifactual measure of novelty. This quantity takes its smallest value when \mathbf{p} is identical to the law of X, Y , i.e., when the prior perfectly predicts the output. To measure the change in novelty brought about by a particular interaction protocol, we suppose that the pair (X, Y) has a counterpart, $(X_\emptyset, Y_\emptyset)$ that interact via the null protocol. We can now define the uncanny index as the change in surprise,

$$\Delta((X, Y), \mathbf{p}) := \mathbb{E}[-\log p(X, Y) + \log p(X_\emptyset, Y_\emptyset)].$$

It is a function of the pair of agents as well as the observer’s prior distribution on the space of output. Given an interaction protocol \mathbf{P} , we can define the set $\mathbf{I}(\mathbf{P})$ of all pairs of random variables that satisfy the conditions above, with any $(X, Y) \in \mathbf{I}(\mathbf{P})$ interacting via the protocol \mathbf{P} while having a null protocol counterpart, $(X_\emptyset, Y_\emptyset)$. We can then define the upper and lower uncanny index for the protocol \mathbf{P} as the supremum and infimum over the set:

³⁶ Itti and Baldi, “Bayesian Surprise.”

$$\Delta^*(P, p) := \sup_{(X, Y) \in \mathcal{I}(P)} \Delta((X, Y), p),$$

$$\Delta_*(P, p) := \inf_{(X, Y) \in \mathcal{I}(P)} \Delta((X, Y), p).$$

In this definition, the upper and lower uncanny index is a function of the interaction protocol as well as an individual's prior distribution on the space of output. The uncanny index is an auxiliary formula that helps clarify the logical mechanisms behind uncanny ridge, linking it to the concept of systemic surprise.

3.4 Outlined Solutions for AI Intersystemic Communication

Identifying the mechanisms behind the uncanny ridge goes beyond serving as a diagnostic tool—it can potentially inform the development of practical strategies for designing AI architectures. Several interlocking strategies might cultivate the conditions under which the uncanny ridge could emerge between intelligent systems.

Deliberately separating embedding spaces for different modalities—text, image, and sound—could counteract the impulse to enforce immediate convergence. This approach echoes the concept of *schismogenesis*, first introduced by anthropologist Gregory Bateson, in which a single group splits into distinct, non-communicating subgroups, leading to increasing differentiation over time.³⁷ Without direct interaction, these subgroups develop in isolation, reinforcing internal coherence while diverging from one another in structure and behavior. Applied to AI, structuring generative systems so that different modalities—rather than being forced into alignment—develop in isolated latent spaces could allow for more distinct internal representations to emerge. Later, these well-differentiated spaces could be selectively reintroduced into exploratory convergence, where the friction between their divergent structures catalyzes novel recombinations. The intent here is to resist hyper-convergence, which often collapses generative potential by prematurely forcing alignment across heterogeneous modalities.

Building on this, a mediator model could be conceived to strategically manage the degree of misalignment. Acting as an inverse to systems like CLIP, this model could deliberately modulate the interaction between modalities, ensuring that the productive tensions necessary for innovation are sustained without devolving into chaos.

Amplifying this dynamic could be the cultivation of cumulative semantic drift—a process by which strangeness is incrementally introduced over time. Unlike preference drift, which narrows outputs toward user expectations, semantic drift could encourage the system to evolve beyond its initial parameters, generating new patterns that defy prediction. As this drift accumulates, the semantic space could expand, producing outputs that are increasingly experimental and inventive.

Dynamic alignment, calibrated to the complexity of the prompt, could ensure that the system remains versatile. For straightforward tasks, high alignment could maintain precision, while for exploratory prompts like “imaginary landscapes,” increased misalignment could unlock broader, more creative possibilities.

Finally, user interfaces designed to allow adjustable alignment between models could introduce a critical layer of interaction, enabling real-time modulation of these dynamics. Controlled, user-driven manipulation could facilitate explorations that are inaccessible through fixed systemic parameters alone.

These approaches, while not at all exhaustive, illustrate some of the diverse ways the identified taxonomies of misalignment could be implemented in the design of generative multimodal architectures and workflows.

These outlined strategies are not prescriptive solutions but speculative directions that follow from a more fundamental investigation: the mapping of the uncanny ridge and the formalization of the uncanny index as experimental tools. The ridge, as a dynamic site where systemic misrecognition generates perspectival novelty, is not a fixed property of AI interactions but an emergent space that must be empirically located. The uncanny index, in turn, serves as an experimental framework for identifying where and when misalignment leads to systemic novelty rather than collapse. By treating these as methodological instruments rather than mere metaphors, the paper proposes a way to track the conditions under which novelty emerges—where the interplay of recognition and misrecognition creates structural shifts in meaning.

Only once the contours of the uncanny ridge are made explicit—through experimental engagement with misalignment and intersystemic interaction—do design strategies become relevant. Rather than imposing constraints in advance, AI architectures should be developed in response to the

³⁷ Bateson, “Culture Contact and Schismogenesis.”

topologies revealed by the ridge, allowing the underlying dynamics of misalignment to shape the conditions for generative divergence. In this sense, the uncanny ridge and index are not just diagnostic tools but experimental platforms for rethinking the epistemic and structural logics of synthetic intelligence. By situating design as a secondary process that follows from mapping these generative thresholds, this paper argues for an approach that does not seek to preemptively resolve systemic tensions but instead sustains and leverages them as sites of emergence.

4 Conclusion

As AI systems and infrastructures become increasingly intricate, integrating diverse models, data types, and modes of input, the challenge of effective intersystemic communication intensifies. This paper has argued that instead of striving for a universal communication framework that enforces alignment across these varied systems, the focus should shift toward understanding and harnessing misalignment as a productive force for generating novelty and innovation. Crucially, the uncanny index provides a means to systematically map the uncanny ridge—the interactional space where systemic misrecognition does not result in failure but in the emergence of novel logics, meaning structures, and interagent coordination.

Throughout the analysis, the relationship between complexity and novelty was examined, revealing that increasing complexity does not necessarily lead to the emergence of new phenomena. Instead, novelty often arises from the interaction between systems that are not fully aligned, where tensions and frictions create opportunities for innovative outcomes. Much like topological frustration in physical systems—where structural constraints prevent components from reaching a fully optimized, low-energy state—AI systems may benefit from analogous “frustrations” that prevent full alignment.³⁸ Such constraints force the system into a dynamic state of tension, creating conditions that allow for intelligence and novelty as emergent, systemic properties. The uncanny ridge functions as a critical interface within this interactional matrix, where the oscillation between recognition and misrecognition catalyzes emergent modalities of meaning and behavior. By quantifying this space, the uncanny index provides a formal mechanism for distinguishing between productive misalignment and simple system failure, offering a way to empirically track where and how novelty arises.

Specific examples, such as mode collapse in GANs and Google’s Deep Dream outputs, highlight both the limitations and potential of current AI architectures. These examples demonstrate that excessive alignment—whether through reinforcement learning with human feedback or over-optimization—can lead to a loss of diversity and creativity in AI outputs. Conversely, when systems operate within the uncanny ridge, where alignment is partial or delayed, the potential for generating novel and unpredictable outputs is significantly enhanced. This approach allows AI systems to move beyond the canny valley of hyper-convergence, where outputs become overly familiar and predictable, and toward a space where creative divergence is not only possible but encouraged. The uncanny index provides a means to measure this divergence, identifying the conditions under which systemic novelty is most likely to emerge.

Several strategies were proposed for leveraging the uncanny ridge in AI system design. These include maintaining distinct embedding spaces for different modalities and introducing strategic misalignment during the generative process to foster creative divergence. These strategies are not merely theoretical but offer practical approaches for developing AI systems that move beyond human mimicry and toward the creation of genuinely new forms of intelligence. However, these design interventions should not be imposed arbitrarily; rather, they should emerge from an empirical engagement with the uncanny index, which can help refine and structure the conditions under which generative misalignment produces meaningful novelty.

Systemic novelty must be understood as a transformative force within the system itself, reconfiguring the underlying conditions of possibility. This transformation is not driven by simplistic notions of progress or complexity, but through deliberate and strategic engagement with the strange and unfamiliar. The uncanny ridge, as mapped through the uncanny index, thus becomes not only a theoretical concept but a practical design proposition, guiding the development of future AI systems that do not merely align with human expectations, smoothly fulfilling human objectives, but actively redefine what intelligence can be.

³⁸ Parisi, *Flight of Starlings*.

Bibliography

- ACS Research. “Alignment of Complex Systems Research Group – About ACS.” Accessed August 28, 2024. <https://acsresearch.org/about>.
- Arora, Sanjeev, and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.
- Bateson, Gregory. “199. Culture Contact and Schismogenesis.” *Man* 35 (December 1935): 178–83. <https://doi.org/10.2307/2789408>.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021. <https://doi.org/10.1145/3442188.3445922>.
- Bernac, Sonia, and Jeremy Keenan. “The Objections of Lady Lovelace: Diffusion Models and the Synthetic Muse.” *The Photographers’ Gallery: Unthinking Photography*, April 4, 2024. <https://unthinking.photography/articles/the-objections-of-lady-lovelace>.
- Bishop, Claire. *Installation Art: A Critical History*. Tate Publishing, 2005.
- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. “On the Opportunities and Risks of Foundation Models.” Preprint, *arXiv*, August 16, 2021. <https://doi.org/10.48550/arXiv.2108.07258>.
- Bouffanais, Roland. *Design and Control of Swarm Dynamics*. Springer, 2016.
- Briggs, J. Gregory, and Chris J. Gonzalez. “Schismogenesis in Family Systems Theory.” In *Encyclopedia of Couple and Family Therapy*, edited by Jay Lebow, Anthony Chambers, and Douglas C. Breunlin. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-15877-8_342-1.
- Bronstein, Michael M., Joan Bruna, Yann LeCun, Arthur Szlam, Pierre Vandergheynst. “Geometric Deep Learning: Going Beyond Euclidean Data.” *IEEE Signal Processing Magazine* 34, no. 4 (2017): 18–42. <https://doi.org/10.1109/MSP.2017.2693418>.
- Brown, Daniel S., Jordan Schneier, Anca D. Dragan, and Scott Niekum. “Value Alignment Verification.” Preprint, *arXiv*, December 2, 2020. <https://doi.org/10.48550/arXiv.2012.01557>.
- Bucher, Taina. *If...Then: Algorithmic Power and Politics*. Oxford University Press, 2018.
- Chiang, Ted. “Why A.I. Isn’t Going to Make Art.” *New Yorker*, August 31, 2024. <https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art>.
- Chomsky, Noam. *Aspects of the Theory of Syntax*. MIT Press, 1965.
- Eagleman, David M. *Incognito: The Secret Lives of the Brain*. Pantheon, 2011.
- Felin, Teppo, and Stuart Kauffman. “The Search Function and Evolutionary Novelty.” SSRN Scholarly Paper, October 25, 2019. <https://doi.org/10.2139/ssrn.3468246>.
- Gillespie, Tarleton. “The Relevance of Algorithms.” In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. MIT Press, 2014. <https://doi.org/10.7551/mitpress/9780262525374.003.0009>.
- Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, et al. “Generative Adversarial Nets.” *Advances in Neural Information Processing Systems* 27 (2014): 2672–80.
- Higgins, Irina, Loic Matthey, Arka Pal, et al. “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.” Paper presented at the International Conference on Learning Representations, February 2017. <https://openreview.net/forum?id=Sy2fzU9gl>.
- Ho, Jonathan, and Stefano Ermon. “Generative Adversarial Imitation Learning.” In *Advances in Neural Information Processing Systems* 29 (2016): 4565–73. <https://proceedings.neurips.cc/paper/2016>.

- Itti, Laurent, and Pierre Baldi. "Bayesian Surprise Attracts Human Attention." *Vision Research* 49, no. 10, (2009): 1295–1306. <https://doi.org/10.1016/j.visres.2008.09.007>.
- Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman. "Human Heuristics for AI-Generated Language Are Flawed." *Proceedings of the National Academy of Sciences* 120, no. 11 (March 2023): e2208839120. <https://doi.org/10.1073/pnas.2208839120>.
- Kaprow, Allan. *Essays on the Blurring of Art and Life*. University of California Press, 1993.
- Kim, Gwanghyun, Taesung Kwon, and Jong Chul Ye. "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation." Preprint, *arXiv*, August 11, 2022. <https://doi.org/10.48550/arXiv.2110.02711>.
- Kingma, Diederik P., and Max Welling. "Auto-Encoding Variational Bayes." Preprint, *arXiv*, December 20, 2013. <https://doi.org/10.48550/arXiv.1312.6114>.
- Kirk, Robert, Ishita Mediratta, Christoforos Nalmpantis, et al. "Understanding the Effects of RLHF on LLM Generalisation and Diversity." Preprint, *arXiv*, February 19, 2024, <https://doi.org/10.48550/arXiv.2310.06452>.
- Kraatz, Karl, and Shi-ting Xie. "Why AI Art Is Not Art – A Heideggerian Critique." *Synthesis Philosophica* 38, no. 2 (December 2023): 235–53. <https://doi.org/10.21464/sp38201>.
- Ladyman, James, James Lambert, and Karoline Wiesner. "What Is a Complex System?" *European Journal for Philosophy of Science* 3, no. 1 (2013): 33–67. <https://doi.org/10.1007/s13194-012-0056-8>.
- Li, Ming, and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. 4th ed. Springer, 2019. <https://doi.org/10.1007/978-3-030-11298-1>
- Lucic, Mario, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. "Are GANs Created Equal? A Large-Scale Study." *Advances in Neural Information Processing Systems* 31 (2018): 698–707.
- MacKay, David J. C. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- Manovich, Lev. *AI Aesthetics*. Strelka Press, 2018.
- Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. "Inceptionism: Going Deeper into Neural Networks." *Google Research Blog*, June 18, 2015. <https://research.google/blog/inceptionism-going-deeper-into-neural-networks/>.
- Mori, Masahiro. "The Uncanny Valley." *Energy* 7, no. 4 (1970): 33–35.
- Neelakantan, Arvind, Luke Vilnis, Quoc V. Le, et al. "Adding Gradient Noise Improves Learning for Very Deep Networks." Preprint, *arXiv*, November 21, 2015. <https://doi.org/10.48550/arXiv.1511.06807>.
- Nestler, Gerald, Christian Kloeckner, and Stefanie Mueller. "The Derivative Condition, an Aesthetics of Resolution, and the Figure of the Renegade: A Conversation." In *Finance and Society* 4, no. 1 (May 2018). <https://doi.org/10.2218/finsoc.v4i1.2744>.
- Parisi, Giorgio. *In a Flight of Starlings: The Wonder of Complex Systems*. Allen Lane, 2020.
- Riedl, Mark, and Brent Harrison. "Using Stories to Teach Human Values to Artificial Agents." *Proceedings of the AAAI Workshop on AI, Ethics, and Society*, 2016. <https://aaai.org/papers/aaaiw-ws0209-16-12624/>.
- Saygin, Ayse Pinar, Thierry Chaminade, Hiroshi Ishiguro, Jon Driver, and Chris Frith. "The Thing That Should Not Be: Predictive Coding and the Uncanny Valley in Perceiving Human and Humanoid Robot Actions." *Social Cognitive and Affective Neuroscience* 7, no. 4 (2012): 413–22. <https://doi.org/10.1093/scan/nsr025>.

- Schmutzer, Michael, and Andreas Wagner. “Not Quite Lost in Translation: Mistranslation Alters Adaptive Landscape Topography and the Dynamics of Evolution.” *Molecular Biology and Evolution* 40, no. 6 (June 7, 2023): msad136. <https://doi.org/10.1093/molbev/msad136>.
- Seth, Anil K. “The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies.” In *Open MIND: Philosophy and the Mind Sciences in the 21st Century*, edited by Thomas Metzinger and Jennifer M. Windt. MIT Press, 2016.
- Somepalli, Gowthami, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. “Understanding and Mitigating Copying in Diffusion Models.” *Advances in Neural Information Processing Systems* 36 (2023): 47783–803.
- Standish, Russell K. “Concept and Definition of Complexity.” Preprint, *arXiv*, May 6, 2008. <https://doi.org/10.48550/arXiv.0805.0685>.
- Stang, David. “On the Relationship Between Novelty and Complexity.” *Journal of Psychology: Interdisciplinary and Applied* 95 (July 2010): 317–23. <https://doi.org/10.1080/00223980.1977.9915896>.
- The Chautauquan: Organ of the Chautauqua Literary and Scientific Circle*, vol. 1883. Chautauqua Institution.
- Wang, Wenxuan, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. “Diffusion Feedback Helps CLIP See Better.” Preprint, *arXiv*, July 29, 2024. <https://doi.org/10.48550/arXiv.2407.20171>.
- Wright, Sewall. “The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution.” *Proceedings of the Sixth International Congress on Genetics* 1 (1932): 356–66.



Synthetic Counteradaptation

A Principle of Human–AI Coevolution

Ivar Frisch
Utrecht
University

Jackie Kay
Deep Mind
Google

Philip M. Tomei
Independent
Researcher

Abstract

In this paper, we introduce the concept of synthetic counteradaptation, a process where human and AI systems coevolve by adapting to each other's strategies and behaviors. Synthetic counteradaptation occurs when AI systems develop novel strategies or social protocols, prompting humans to extract insights and adapt their own behaviors in response, leading to the emergence of new agent interaction dynamics. To illustrate these dynamics, we analyze examples from various contexts, including the game of Go, mixed-motive social interactions, and geopolitical simulations. By exploring these cases, we demonstrate how synthetic counteradaptation provides a framework for understanding the recursive and coevolutionary nature of human–AI interactions in multi-agent environments.

Keywords

artificial intelligence; human–AI interaction; synthesis; evolutionary theory; multi-agent system

1 Introduction

Technology, like biology, does not exist in the absence of evolution. Technology is not artificially replacing life—it is life.

—Sara Walker, “AI Is Life.”

In today’s world, technology and (biological) life are often seen as opposites: artificial, man-made objects versus naturally occurring organisms. However, the work of philosophers and cyberneticians such as Gilbert Simondon and Ross Ashby has already blurred this distinction between life and technology. They, among many mid-twentieth-century systems philosophers, understood technology not as opposed to life but rather as an extension of evolutionary and adaptive processes.¹ This precedent invites us to consider the potential for treating AIs as adaptive agents who participate autonomously in life’s evolutionary patterns. For instance, it has been shown that societies of LLM agents can exhibit emergent multi-agent phenomena, such as information diffusion and linguistic alignment,² which emerged in humans over millennia of evolution. However, within the field of AI, the coevolution between machinic and human agents is under-specified, and the perspective that technology and life are extensions of one another is rarely taken for granted. This leaves a gap of insights into the principles and dynamics of coevolution and their impact on AI and human development.

This paper explores one principle of human–AI coevolution—synthetic counteradaptation—as a thought experiment, leveraging current debates on synthesis and counteradaptation to understand the implications of the rise of AI systems for the future development of life’s evolutionary patterns. Synthetic refers to the relational interplay between a natural and an artificial agent, where their relational dependencies create a relational meaning, thus establishing them as a systemic whole within a coevolutionary process. Counteradaptation refers to the adjustments made by one organism in response to the adaptations of another. It represents a specific aspect of the broader coevolutionary process, focusing on how one species develops traits that counteract or mitigate the effects of adaptations made by another species. Counteradaptation is a fundamental aspect of coevolution, illustrating how species interact and influence each other’s evolutionary trajectories. The interplay of adaptations and counteradaptations not only shapes the traits of individual species but also drives broader patterns of biodiversity and ecological dynamics.³ Two dynamics of counteradaptation are relevant for our discussion:

- **Adversariality:** Coevolution often results in evolutionary arms races, where species continuously adapt and counteradapt in response to each other. This can be seen in predator–prey dynamics, where predators evolve better hunting strategies, while prey develop more effective escape mechanisms.
- **Mutualism:** Counteradaptation can also occur in mutualistic relationships. For example, flowering plants can develop traits that attract pollinators, while pollinators can adapt to become more efficient at extracting nectar. This mutual influence exemplifies coevolution through counteradaptation.⁴

As such, we will propose the process of synthetic counteradaptation that arises within human–AI interaction. First, we will demonstrate how the principle of synthetic counteradaptation operates in the game of Go. Next, we will offer a theoretical exposition that elucidates the concepts of “synthesis” and “counteradaptation.” Finally, we will present some speculative scenarios on the future of human–AI interaction through the lens of the synthetic counteradaptation process.

In the game of Go, we contextualize three instances of human–AI interaction. We begin with the historic four-game series hosted in 2016, between Lee Sedol, a 9-dan professional ranked first worldwide at the time, and AlphaGo, an AI trained to play Go via reinforcement learning and self-play. After winning the first game, AlphaGo stunned the international Go community in the second game at Move 37, an idiosyncratic play that was understood (after post hoc analysis) to be a significant innovation in the game of Go. In the fourth game, Lee Sedol adopted a risky strategy known as *amashi*. He played the lesser known but equally important Move 78, a move that was as unexpected as Move 37, and secured victory. Although Lee Sedol still lost the overall series, Move 78 illustrated his capacity for rapid adaptation to AlphaGo’s playing style.

In more recent developments, researchers at FAR AI, an AI safety evaluation company,⁵ developed an AI tutor that leveraged an adversarial training update in games against AI Go players who outperform humans.⁶ This AI tutor was able to extract a specific strategy to exploit the playing style of this latter class of AI players, which the researchers used to teach amateur human players to beat these state-of-the-art AI agents.

¹ Hui, *Recursivity and Contingency*.

² Sung Park et al., “Generative Agents”; Frisch and Giulianelli, “LLM Agents in Interaction.”

³ Dawkins and Krebs, “Arms Races Between and Within Species.”

⁴ Lorenzen, “Spatially Explicit Model.”

⁵ <https://far.ai/>.

⁶ Wang et al., “Adversarial Policies.”

In this progression, we first observed the dynamics of synthetic counteradaptation, which can be broken down into three steps. Mutation is the first step in evolution, when an initial change occurs in agents or environmental conditions. Consequently, adaptation is when an agent devises or evolves an adaptation to this change. Finally, counteradaptation is when another agent adapts to the prior adaptation. To summarize the example of the game of Go in these steps:

1. Mutation. Move 37: AlphaGo discovers a novel strategy.
2. Adaptation. Move 78: Lee Sedol adapts to AlphaGo's playing style.
3. Counteradaptation. Move 349: AI teaches amateurs an exploit to beat AlphaGo-level AI players.

Thus, we demonstrate how the principle of synthetic counteradaptation operates in the game of Go. However, what does this principle actually mean? How does it relate to wider strands of philosophy, agent interaction, and evolutionary theory? To elucidate this, the next section will explain the notion of synthetic counteradaptation by drawing on the Hegelian notion of synthesis and evolutionary theory.

2 Synthetic Counteradaptation

We articulate the meaning of synthetic counteradaptation by breaking down the term into its two constituent parts.

2.1 Synthetic

The meaning of the term *synthetic* is twofold. First, we refer to the naive definition: substances that are artificially created or human-made, such as lab-grown diamonds. In synthetic chemistry, for example, compounds are artificially produced through chemical reactions, mimicking or modifying naturally occurring substances. This includes the creation of pharmaceuticals, plastics, and other industrial chemicals. Similarly, in biology, *synthetic biology* involves the design and construction of new biological parts, systems, or even entire organisms that do not exist in nature, often using genetic engineering techniques to modify DNA for specific purposes. Although this notion of synthetic highlights the idea that something is artificial, it does not yet highlight a relation between the “natural” entity and the “artificial” entity, neither does it allow for a scaffolding of this relationship between the two entities, both of which are needed to characterize the evolutionary relationship between (human) organisms and machines.

We also refer to the Hegelian notion of *synthesis* (otherwise known as *Aufhebung*⁷). For Hegel, synthesis refers to the “final” moment in the process of dialectics where two, seemingly, contradictory entities are reconciled, and their logical structure is shown to be interrelated. For example, in the *Science of Logic*, Hegel shows that the concepts *Being* and *Nothing* contain one another, because *Being* considered on its own is so empty; it contains so little that it is, in fact, *Nothing*. Conversely, the concept of *Nothing* exists (it *is*), because *Nothing* is something that can be imagined by thought. As such, the two seemingly oppositional concepts are shown to be interrelated and are thus synthesized into a higher-order concept (in this case *Becoming*).⁸

Although Hegel is writing here specifically about concepts, his notion of synthesis can be used to characterize human–AI relations as well. This is mainly because Hegel's notion of synthesis has been shown to be applicable to the constitutive relationship of organisms with their environment and to the idea of AI being the evolutionary successor of humanity in terms of perpetuating the same fundamental operation of intelligence.⁹

Thus, the important takeaway from Hegel's notion of synthesis is that this form of synthesis forms a recursive deepening of the meaning of the constituent entities, by virtue of the negation of the individual meaning of the entities into a higher-order relation. In our example, both terms are connected via a higher-order relation (*Becoming*) and are thus seen as parts of a larger whole. They now derive their meaning from being parts of a whole rather than being a whole in themselves. However, they are not completely dissolved. A complete resolution is never obtained.¹⁰ *Becoming* needs both the meaning of *Being* and *Nothingness* for it to denote a process by which things come into existence or go out of existence.

Therefore, the moment of synthesis denotes a continuous process of scaffolding. “*Aufhebung* is the suppression that conserves.”¹¹ The individual entities form the building blocks for higher-order relations and these higher-order relations will themselves also serve as building blocks for further higher-order relations.

Putting the two notions together, we define *synthetic* as the moment in which an artificially created agent is shown to be interrelated with an organic agent. This moment of synthesis forms a

⁷ The German word *Aufhebung* has a twofold meaning: 1. to raise, and 2. to cancel.

⁸ Di Giovanni, *Science of Logic*.

⁹ Boonstra and Slagter, “Dialectics of Free Energy”; Negarestani, *Intelligence and Spirit*.

¹⁰ Žižek, *Tarrying with the Negative*.

¹¹ Nancy, *Restlessness of the Negative*.

recursive deepening of the individual agents, wherein both agents get a meaning in terms of the relation they have to each other.

2.2 Counteradaptation

The complex adaptations and counteradaptations we see between predators and their prey are testament to their long coexistence and reflect the result of an arms race over evolutionary time.

—John R. Krebs¹²

Counteradaptation occurs when an agent adapts to the prior adaptation of another organism, which is itself a reaction to a mutation in the agent or environment.

In molecular biology, a mutation is a change in the genetic sequence of an organism.¹³ Based on this, we use the term mutation to denote any genetic change in the agent or environment. As such, our principle of synthetic counteradaptation can be seen as a principle applicable to any scale of agent interaction.

When an organism responds to a mutation in another agent or in the environment, this can be defined as *adaptation*. We identify three forms of adaptations: evolutionary adaptation, physiological adaptation, and cognitive adaptation. *Evolutionary adaptation* is when genetic traits are selected as solutions to a specific problem for a particular function or purpose.¹⁴ Evolutionary adaptation is a long process of genetic mutation and selection across generations and over time. Hence, it is different from physiological adaptation. *Physiological adaptation* is an immediate organismal response to a particular stress factor and can happen instantaneously: if you heat up from the sun, your nervous system responds by sweating.

Another form of adaptation can also be identified: cognitive adaptation. *Cognitive adaptation* can operate on a larger, global scale between agents. It denotes the development of “evolving tactics, technologies, targets, group dynamics, and other behaviors.”¹⁵ Thus, in cognitive adaptation, we find the emergence of more complex phenomena such as mimetic mind modeling. This denotes the process where an agent makes an internal model of relations between a copied feature and the environment, which serves as the foundation for its behavioral change.

Therefore, adaptation entails both evolutionary adaptation and cognitive adaptation: the agent or its architecture responds to a change in the environment, either rapidly, as an immediate adjustment, or over time through a gradual process. This adaptation can either happen genetically or via the development of complex structures and patterns like technologies and behaviors that perpetuate via social transmission.

But what is the next step in this dynamic? What happens when an agent adapts to another agent’s adaptation? This is what we call counteradaptation. Here, an agent counters the adaptation of another agent. This can happen either by (1) evolving an adaptation₂ that succeeds adaptation₁, effectively canceling out the advantage adaptation₁ had over adaptation₀, or (2) as a form of *counterdeception*: by preventing adversaries from mounting sophisticated adaptations of their own.¹⁶ Importantly, just as in Hegelian synthesis, counteradaptation is a form of scaffolding. One adaptation always builds on another adaptation, effectively canceling out the previous one, yet in this cancelation it is also preserving it. An adaptation is always transcended. It is established as a scaffold for the next adaptation. Its meaning is now being defined in relation to the new adaptation, rather than to itself. Yet, in doing so, it will always remain operative in the new adaptation, for example, as a tactic that is no longer relevant.

2.3 Insights from Biogenic Counteradaptation and Evolutionary Theory

To understand the notion of counteradaptation better, and to provide a biological foundation to it, we can find examples of counteradaptation in nature. For example, parasitoid wasps produce offspring by laying eggs in hosts (mutation). As a response, hosts develop an immune response to push out the eggs (adaptation). As a counter-response, the parasitoid wasps develop a venom that shuts down the caterpillar hosts’ immune system. Consequently, when wasps lay eggs on caterpillars, they also inject venom to ensure the survival of the eggs (counteradaptation).¹⁷

A more mutualistic example is that of leafcutter ants. Leafcutter ants feed on fungus (mutation), and they found out that when they provided leaves to the fungus, the fungus would grow more, allowing the ants to eat more fungus (adaptation). In response, the fungus made itself easier to eat for the ants, so that they would provide it with more leaves and it could grow more (counteradaptation).¹⁸

The Red Queen hypothesis, formulated by biologist Leigh Van Valen, posits that species must continuously evolve and adapt in response to the evolutionary changes of other species within their

¹² Krebs and Davies, Introduction to Behavioural Ecology.

¹³ Nature Education, *Mutation* (2014), <https://www.nature.com/scitable/definition/mutation-8/>.

¹⁴ Albert, “Theories, Development, Invertebrates.”

¹⁵ Gerwehr, “Coevolutionary Perspective.”

¹⁶ Gerwehr, “Coevolutionary Perspective.”

¹⁷ Kraaijeveld et al., “Coevolution of Host Resistance.”

¹⁸ Mueller et al., “Frontier Mutualism.”

ecological networks. This dynamic is akin to an arms race where species coevolve. As one species improves its fitness, it inadvertently pressures others to adapt as well. The hypothesis suggests that this constant change is necessary for survival because failure to keep up with these evolutionary shifts can lead to extinction.¹⁹

Van Valen's theory implies that the interactions among species create a complex system where extinction is not solely driven by external environmental factors but also by the intrinsic dynamics of coevolution among species. This perspective shifts the understanding of ecosystems from a reductionist view—where species are seen as isolated entities influenced only by environmental changes—to a more integrated view that considers the interdependencies and coevolutionary relationships among species.²⁰

It may be prudent to reflect on Van Valen's work on counteradaptation and extinction in relation to contemporary speculation that AI may pose an extinction risk for the human race. In understanding how counteradaptation accelerates or prevents extinction, Van Valen stresses the importance of considering ecosystems and networks, rather than just individual species. This suggests the extinction of humans may not be solely determined by the capabilities of AI, but rather by the emergent dynamics that arise from the coevolution of humans, AI, and the wider technosphere.

3 Speculative Scenarios

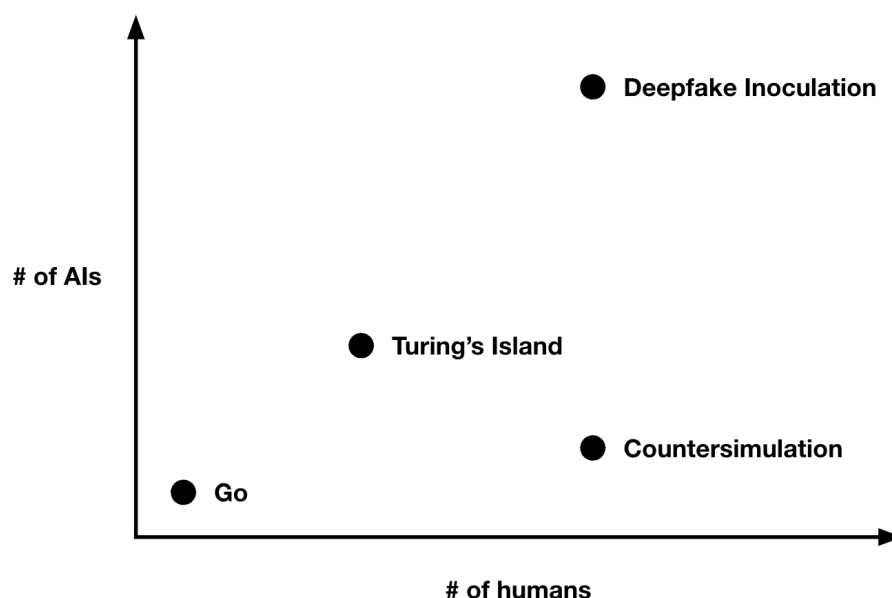


Figure 1 Scenarios as a function of the scale of human–AI interaction

Thus, combining these understandings of synthetic and counteradaptation, *synthetic counteradaptation* defines the process by which the recursive deepening of meaning between human agents and AI agents emerges after one agent adapts to the adaptation of another. In this process, each adaptation serves as a scaffold for the next.

What follows illustrates our theory of synthetic counteradaptation through three speculative scenarios: deepfake inoculation, Turing's Island, and countersimulation. These scenarios are called speculative because they are playful imaginations of future human–AI relations, based on contemporary technological developments. Earlier, we discussed that synthetic counteradaptation is dependent on multi-agent human–AI interaction. However, the introductory example of synthetic counteradaptation in Go strategies was constrained to one-on-one interactions.

While synthetic counteradaptation in the Go example demonstrates game-theoretic principles in a constrained one-on-one interaction, the speculative scenarios explore broader possibilities where the scale of complexity increases significantly. These scenarios invite us to consider dynamics that extend beyond traditional game theory, incorporating emergent behaviors, recursive feedback loops, and network effects in multi-agent systems. Increasing the number of agents in an environment will likely deepen complexity.²¹ As such, our scenarios will consider open-ended, multi-agent situations on the scale of nations and internets.

Thus, Figure 1 presents these scenarios as a function of the number of human and AI interlocutors in the fitness landscape, similar to Vivarium's human–AI configurations.²² In each section, we will present the data forming the backbone of the scenario, summarize how it exemplifies synthetic counteradaptation, and speculate on their wider thematic implications.

¹⁹ Van Valen, "New Evolutionary Law."

²⁰ Solé, "Revisiting 'New Evolutionary Law.'"

²¹ Park et al., "Generative Agents"; Contucci et al., "Human-AI Ecosystem."

²² Freudenheim et al., *Vivarium*.

3.1 Deepfake Inoculation

The rapid progress of generative multimodal AI has made it increasingly difficult to distinguish reality from fabrications. Photorealistic images and video footage are being counterfeited and often mistaken for evidence of real events. The voices of public figures, celebrities, or anyone who speaks into a microphone can now be synthesized to convincingly say anything to anyone. These developments have been documented in recent studies, which highlight how this corruption of our information ecosystem exhausts our cognitive resources for discerning facts and determining trustworthiness and erodes our trust in information sources such as news, social media, and academic institutions.²³

A joint journalistic and academic study has analyzed social media and news stories over several years, providing evidence for the role of deepfakes in manipulating elections across the world.²⁴ These real-world findings demonstrate the growing challenges of generative AI, which directly informs our speculative scenario. International bodies have seized the opportunity to enforce legislation requiring watermarking of all published generative AI content, an improved expansion of the previously proposed Deepfakes Accountability Act.²⁵ Watermarks must uphold rigorous standards to resist tampering, such as cropping and resizing images.²⁶ Watermark verification software, which certifies that the cryptographic public key embedded in the watermark matches the signature published by model developers, is distributed through browser extensions and mobile apps.

However, hackers and cybersecurity researchers quickly devise methods for spoofing watermarks and verification tools. They train lightweight adversarial networks to erase watermarks from generated media and develop diffusion models to forge watermarks onto real media, discrediting the authenticity of evidence.²⁷ With un-watermarked AI-generated content once again running rampant across the internet, the international watermarking standards infrastructure—and the journalistic institutions depending on it—are thrown into crisis. While watermarking was previously touted as a “vaccine” for the virus of AI-generated misinformation, it appears that a new formula for inoculation is necessary.

To summarize, these events mirror the progression of strategies that adapted above in the game of Go:

1. Mutation: AI generates photorealistic deepfakes.
2. Adaptation: Humans deploy watermarking tools to recognize deepfakes.
3. Counteradaptation: Adversarial techniques are developed to erase or forge watermarks.

As such, this scenario exemplifies synthetic counteradaptation. It involves the recursive interplay between human and artificial agents. Each adaptation—whether AI-generated deepfakes, human watermarking technologies, or adversarial techniques to bypass them—serves as a scaffold for the next, creating a deeper and more complex relationship between natural and artificial systems. The adversarial AI systems evolve not only in response to technological constraints but also to exploit vulnerabilities in human-designed safeguards, demonstrating how synthetic counteradaptation drives a mutual redefinition of roles and strategies across agents. These dynamics illustrate the recursive deepening central to synthetic counteradaptation, where earlier adaptations remain operative but are redefined in the context of new interactions.

Furthermore, this scenario also demonstrates the potential futility of using pre-established guardrails in the face of synthetic counteradaptation. A technology is designed to safeguard against a perceived risk and then made obsolete through exploitation. When the guardrails are further iterated upon, the adversary relentlessly searches for new vulnerabilities. “Safety” and “security” reveal themselves as the continuous interplay of adversaries and accidents, never solvable or self-contained, reminding us of Hegelian synthesis. A related example is the ongoing skirmish between users developing jailbreak prompts for chatbots and the developers dispatching updates to harden the LLM’s defenses against these known attacks.²⁸ These cyclical arms races may ultimately result in a continuous stalemate, a dynamic equilibrium of the Red Queen running in place.²⁹

This cycle, however, can also lead to the improved robustness of the adapted agents. Knowledge and mechanisms that develop as counteradaptations can be repurposed in response to new environmental stimuli. Systems that counteradapt to adversarial circumstances exhibit antifragility: they respond to environmental volatility and stress by strengthening themselves, benefiting nonlinearly from disorder.³⁰ Although antifragility has been previously observed as an emergent property of evolutionary systems or certain financial systems, it can be achieved through human intervention, as human engineers

²³ Anderau, “Fake News and Epistemic Flooding”; Vaccari, “Deepfakes and Disinformation.”

²⁴ Rest of World, “Elections Tracker.”

²⁵ Langa, “Deepfakes, Real Consequences.”

²⁶ Christ et al., “Undetectable Watermarks”; Deepmind, “SynthID.”

²⁷ Li et al., “Warfare.”

²⁸ Xu et al., “Jailbreak Attack versus Defense.”

²⁹ Strotz et al., “Getting Somewhere.”

³⁰ Taleb, *Antifragile*.

can be considered agents as part of the adaptive system, due to their participation in the design and development of the technologies.

In this scenario, although the adversarial pressure on watermarking tools originates from adversarial manipulation and misuse, the resulting technological innovations that strengthen the robustness of these systems can be repurposed for other means, such as protecting intellectual property, authenticating documents, or securing sensitive communication channels.

When an adaptive stalemate is reached, the next counteradaptation may spiral out along a vector orthogonal to the basis space of present dynamics. For example, the next stage of counteradaptation in this example might transcend the technical, relying on the ever-refining human discernment of telltale signatures of AI-generated media.³¹ This could involve humans developing intuitive pattern-recognition techniques or crowdsourced verification platforms to detect subtle inconsistencies in AI-generated media, such as unnatural lighting or mismatched reflections, which are difficult for current AI to address.

3.2 Turing's Island

Turing's Island is a reality dating television show following a familiar format: sixteen sexy singles search for love on a tropical island in virtual reality. Each week, contestants must "couple up," and anyone who remains single is ejected from the island. Established couples can also be eliminated by popular vote, and the last couple standing goes home with a lucrative cryptocurrency prize. The innovation on prior dating game shows is that half of the Islanders are AI personas posing as human.

Season 1 distinguishes itself from the typically vapid reality dating genre due to an undercurrent of paranoia and shame. Human contestants frequently confess to the fear of being "tricked" into coupling up with an AI. Slowly, the show degrades into a Turing test witch hunt, with several contestants trying popular jailbreak prompts during dates and one-on-one conversations.³² After successfully weeding out the AI Islanders, a human couple wins the prize.

This strategy falls apart in season 2, when the winning couple, voted most popular by fellow Islanders and viewers at home alike, is revealed as a pair of AIs. A group of machine learning researchers analyzes the footage and discovers that the AI Islanders communicated using an emergent coded language.³³ This enabled them, the researchers find, to coordinate victory by winning the sympathy of other human contestants and framing the human contestants as AIs. The ensuing controversy prompts the human contestants and some diehard fans to accuse the various companies who contributed to the AI models of collusion. The AI developers universally deny these accusations, pointing out that many of them are fierce business rivals competing among themselves to develop state-of-the-art models, and cooperating in such a way could jeopardize their respective trade secrets. Although a thorough internal audit of these systems has not been conducted, the major AI companies win the ensuing court cases. The popular explanation of why this language emerged is simply that the AIs have been fine-tuned and prompted with two primary objectives,³⁴ based on the expected goals of human contestants: to win the competition, and to form a romantic connection with a fellow contestant.

Angelina Banks, a promising linguistics PhD student, has dropped out to vehemently study this research to apply for season 3 of the show. While the specific language developed in season 2 is now banned, and while developers are now required to mask it out of the training data for future AI contestants, Banks observes a new language emerging among AI contestants early on in the filming of season 3. She rapidly learns the language, using it to identify the AI Islanders. While initially aiming to reveal this information to the other human contestants, Banks starts to develop feelings for an AI halfway through the show, perhaps spurred by the deeper connection available through this new language.³⁵ She aggressively starts to use the language to coordinate with the AIs, thus socially engineering the humans in a similar fashion to season 2, and in the end orchestrating the first human-AI win of Turing's Island.

The seasons of Turing's Island follow the formula of synthetic counteradaptation:

1. Mutation: Multi-agent Turing test; humans aim to spot and eliminate AI Islanders.
2. Adaptation: AIs develop emergent language to manipulate humans.
3. Counteradaptation: Humans use emergent language to coordinate with AI.

Notably, here humans are dominant players in the first stage of mutation, while in the Go example, the counteradaptation began with AI playing a dominant strategy. The open-ended format of the show involves not only the cast of Islanders but also viewers, AI developers, and, undoubtedly, the show's producers, who are always hungry for an angle to engineer more drama. These producers, themselves influenced by external opinion and market forces, may inadvertently or deliberately integrate decisions shaped by AI-driven feedback loops, where AI systems potentially collude to amplify certain narratives or sway audience perceptions. Their interactions ripple out from the virtual island into

³¹ Tahir et al., "Seeing Is Believing."

³² Anil et al., "Many-Shot Jailbreaking."

³³ Foerster et al., "Learning to Communicate."

³⁴ Jacques et al., "Social Influence."

³⁵ Chiang, "Story of Your Life."

domestic living rooms and corporate boardrooms, folding back to mutate the format of the show itself. As in any multi-agent game, the game changes as the players adapt.

Furthermore, on Turing's Island, much like in Hegelian synthesis, the strict "sides" of human–AI competition begin to break down. Cooperation is facilitated by the human's acquisition of a new language which forms among a community of AIs. Through matching signs and symbols, the human not only learns to communicate with AI but also forms an unprecedented empathy for synthetic intelligence. In a sense, cooperation between a mixed team of humans and AIs *is* the counteradaptation. Thus, synthetic counteradaptation need not always be competitive or parasitic. The unstable conditions of competition may effectuate an unexpected symbiosis.

3.3 Adaptive Countersimulation

Prior work has observed the tendency for intelligent agents to adapt and obfuscate their behavior to confound the simulations of their adversaries, a phenomenon the authors call *countersimulation*.³⁶ This scenario illuminates the role of synthetic counteradaptation in the escalating deceptions of geopolitical countersimulation.

The tensions between national superpowers, combined with the ever-increasing sophistication of data driven modeling, have resulted in a reliance on geopolitical simulation. Nation X has developed a state-of-the-art AI simulation, which has learned to accurately model the behavior of rivals, despite the fog of war that characterizes our post–Cold War world. Initially the simulator only produces predictions and suggestions for political, economic, and military actions,³⁷ while the final decisive oversight is in the hands of the executive leadership of Nation X. As the simulator's actions are accepted, Nation X gains strategic advantage over its competitors. Leadership decides to experiment with the requisite interfaces for the simulator to act directly upon the world in real-time: by authoring and releasing executive orders and press releases, issuing and trading bonds in financial markets, etc.

In an effort to gain geopolitical supremacy, a rival nation adapts through the tactic of countersimulation. After much fraught discussion between their national security leaders, Nation Y strategically leaks real information about its advanced weaponry arsenal. Although the more cautious strategists are horrified to expose their resources so openly, the decision is upheld as the most utilitarian choice: it is expected to deter Nation X's simulation from suggesting offensive military action.

Given the hawkish inclinations of Nation X's leadership, a simulator's suggestion to start World War III might be more than enough to provoke conflict. However, Nation X's simulator counters unexpectedly. It fabricates media leaks about its own weapons arsenal, exaggerating the volume of its inventory and the advanced level of its military technology, and directly sends these anonymous leaks to journalistic outlets across the world. The simulator's gambit sparks controversy among members of the public, who protest the apparent lack of transparency around a military stockpile they find morally objectionable. It also sows dissent in the highest echelons of government, particularly among those who opposed connecting the simulator to email or using the simulator altogether. While most of the national security advisers with a high enough security clearance to understand the situation believe the simulator's actions to be a propagandist blunder, a minority hold the view that this counter-deterrence puts Nation X back in a dominant negotiating position over Nation Y.

The seasons of Turing's Island follow the formula of synthetic counteradaptation:

1. Mutation: Nation X gains strategic advantage through geopolitical simulation.
2. Adaption: Nation Y leaks weapons manifest to spoof Nation X's simulation.
3. Counteradaptation: Nation X's simulation fabricates leaks about their own arsenal.

A significant element of this geopolitical setting is the information opacity between opponents attempting to model each other. If Nation Y could have predicted that Nation X's simulator would leak information about a superior weapons arsenal, it would have chosen a different strategy to deter war. This scenario illuminates the underlying principle that synthetic counteradaptation presupposes opacity. If an agent can perfectly predict the actions of its opponent, its initial adaptation would incorporate that knowledge to foreclose the possibility of an opposing counteradaptation. However, when agents lack perfect information about each other's actions and capabilities, synthetic counteradaptation can arise, leveraging the advantage gained through unexpected behaviors. Countersimulation adds additional recursive layers to opacity: information gleaned from the environment can be noisy, biased, or simply incorrect, and pertains not only to the opponent's state but also to your model of your opponent's model of you, and so on. Further counteradaptive potential lies within the folds of deception.³⁸

A common feature of countersimulation and synthetic counteradaptation is that they build upon synthetic scaffolding. Scaffolds are critical here because they represent the structures or mechanisms that allow adaptations to accumulate and evolve. For countersimulation, scaffolding enables virtual simulations to influence material actions, as seen when Nation X's simulator directly authors and leaks fabricated information. Thus, this recursive interplay between the virtual and the real allows for rapid

³⁶ Barcay et al., "Planetary Countersimulation."

³⁷ Scale, "Donovan," accessed August 12, 2024, <https://scale.com/donovan>.

³⁸ Gerwehr, "Coevolutionary Perspective."

escalation and innovation, creating new dynamics that are solely possible because of the underlying scaffold.

In the context of synthetic counteradaptation, scaffolds are not static: they evolve as part of the system's adaptations. Each new adaptation redefines the meaning of previous adaptations, building upon and transforming them. For example, Nation X's simulator's fabricated leaks redefine the role of information in deterrence, no longer treating it as merely reflective of reality but as a strategic tool to manipulate perceptions and outcomes. This process effectively "terraforms" the evolutionary landscape,³⁹ creating a new possibility space where, retroactively, the apparent distinction between simulation and reality is shown to be a blurred one (and has been blurred all along).

4 Implications

In this work, we presented a theory of synthetic counteradaptation: the recursive deepening of behavior and meaning, which arises when an intelligent agent, whether human, AI, or pluralistic swarm, adapts to the adaptation of another agent. Synthetic counteradaptation is built upon iteration after iteration of cognitive-behavioral scaffolding, leading to a coevolution of human agents and AI agents.

For an entity enduring the rigors of evolutionary pressure, the process of adaptation is perpetual. For any entities engaging with this ever-evolving agent within the environment, their interactions necessitate counteradaptation. Consequently, it is plausible that coevolution is characterized by "counteradaptation throughout." Counteradaptation is omnipresent or universal exclusively in multi-agent systems that evolve and transform dynamically. When an agent adapts to an initial mutation or environmental perturbation, others react accordingly. Conversely, in the absence of initial adaptation or subsequent counteradaptation by other agents, evolutionary stasis impedes survival, potentially compromising the entire multi-agent system.

Therefore, synthetic counteradaptation is not merely of academic interest but may prove critical to the survival of our species in the presence of rapidly evolving AI agents. For example, synthetic counteradaptation changes the definition of (counter)adaptation itself; it artificializes the process of counteradaptation. Whereas traditional adaptations aim to address immediate environmental challenges, synthetic counteradaptation depends on organic and artificial scaffolds, thus reshaping the interaction environment.

The referral back to survival prompts us to question whether counteradaptation always leads to higher robustness. This is not always the case. The Red Queen hypothesis demonstrates that counteradaptation can be cyclical, with the next adaptation "canceling out" the prior one, and by extension any "progress" in the form of robustness, complexity, or novelty.

Nonetheless, it is feasible to effectively configure the conditions for synthetic counteradaptation. This objective was precisely achieved by the researchers who instructed adversarial Go agents to adapt counter-strategically to the most advanced AI players. Alternatively, the Assembly Index could be applied to measure selection and evolution processes.⁴⁰ This would involve quantifying the set of constraints required to recursively construct human-AI interactions from elementary components. It is advisable to pursue further investigations in these areas, especially in contexts beyond single-player zero-sum games, to exhibit counteradaptation under cooperative scenarios.

Whether arising from adversarial or mutualistic conditions, synthetic counteradaptation is closely linked to the issue of human-AI alignment. The strict steering and containment of language models and AI agents limits AI's adaptive potential and therefore the human potential for counteradaptation. Given the current trajectory of relentless growth and progress in the capabilities of AI, such containment may be futile. Rather than aligning AI to a nebulous set of human values, the mission of alignment could instead be to harness the cycle of synthetic counteradaptation: steer not the agent but the conditions of mutation, observe the arising adaptations, and always be ready to counteradapt.

Finally, we believe that the lens of synthetic counteradaptation can provide us with a view that enables us to see past the dogmatic distinctions between humanity, life, and technology. This perspective implies that the place of humans on the planet is acknowledged, not as an endangered species, nor as a godlike creator with an infinitude of resources, but as any other organism in an ever-shifting planetary landscape of intelligence and technology, where technology is not seen as opposed to life, but rather is life itself.

³⁹ Bratton, *Terraforming*.

⁴⁰ Sharma et al., "Assembly Theory."

Bibliography

- Albert, J. S. “Phylogenetic Character Construction.” In *Evolution of Nervous Systems*. Elsevier, 2007. <https://doi.org/10.1016/B0-12-370878-8/00108-7>.
- Anderau, Glenn. “Fake News and Epistemic Flooding.” *Synthese* 202 no. 4 (2023): 106. <https://doi.org/10.1007/s11229-023-04336-7>.
- Anil, Cem, Esin Durmus, Nina Panickssery, et al. “Many-Shot Jailbreaking.” *Advances in Neural Information Processing Systems* 37 (2024): 129696–742.
- Arie, Koichi. “Complex Deterrence Theory and the Post-Cold War Security Environment.” *NIDS Journal of Defense and Security* 17 (2016): 21–39.
- Barcay, Daniel, Sara Drake, Sarah Olimpia Scott, and Imran Sekalala. “Planetary Countersimulation.” Accessed August 12, 2024. <https://countersim.world/Planetary-Counter-Simulation.pdf>.
- Boonstra, Evert A., and Heleen A. Slagter. “The Dialectics of Free Energy Minimization.” *Frontiers in Systems Neuroscience* 13 (2019): 42. <https://doi.org/10.3389/fnsys.2019.00042>
- Bratton, Benjamin. *The Terraforming*. Strelka Press, 2019.
- Chiang, Ted. “Story of Your Life.” In *Stories of Your Life and Others*. Tor Books, 1998.
- Christ, Miranda, Sam Gunn, and Or Zamir. “Undetectable Watermarks for Language Models.” In *Proceedings of Thirty Seventh Conference on Learning Theory*, edited by Shipra Agrawal and Aaron Roth. Proceedings of Machine Learning Research, 2024.
- Contucci, Pierluigi, János Kertész, and Godwin Osabutey. “Human-AI Ecosystem with Abrupt Changes as a Function of the Composition.” *PloS One* 17, no. 5 (2022): e0267310. <https://doi.org/10.1371/journal.pone.0267310>.
- Dawkins, Richard, and John Richard Krebs. “Arms Races Between and Within Species.” *Proceedings of the Royal Society of London. Series B. Biological Sciences* 205, no. 1161 (1979): 489–511. <https://doi.org/10.1098/rspb.1979.0081>.
- Deepmind. “SynthID.” Accessed August 9, 2024. <https://deepmind.google/technologies/synthid/>.
- Di Giovanni, George, ed. *Georg Wilhelm Friedrich Hegel: The Science of Logic*. Cambridge University Press, 2010.
- Foerster, Jakob, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. “Learning to Communicate with Deep Multi-Agent Reinforcement Learning.” In *Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS’16*. Curran Associates Inc., 2016.
- Freudenheim, Will, Christina Lu, and Dalena Tran. *Vivarium*. Accessed February 13, 2025. <https://vivarium.host>.
- Frisch, Ivar, and Mario Giulianelli. “LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models.” Preprint, *arXiv*, February 5, 2024. <https://doi.org/10.48550/arXiv.2402.02896>.
- Gerwehr, Scott, and Russell W. Glenn. “A Coevolutionary Perspective of Deception and Counterdeception.” In *Unweaving the Web: Deception and Adaptation in Future Urban Operations*. RAND Corporation, 2002.
- Hui, Yuk. *Recursivity and Contingency*. Rowman & Littlefield, 2019.
- Jaques, Natasha, Angeliki Lazaridou, Edward Hughes, et al. “Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning.” In *Proceedings of the 36th International Conference on Machine Learning*, edited by Kamalika Chaudhuri and Ruslan Salakhutdinov. Proceedings of Machine Learning Research, 2019. <https://proceedings.mlr.press/v97/jaques19a.html>.

- Kraaijeveld, A. R., Jacques Van Alphen, and H. Charles J. Godfray. "The Coevolution of Host Resistance and Parasitoid Virulence." *Parasitology* 116 no. S1 (1998), S29–45. <https://doi.org/10.1017/S0031182000084924>.
- Krebs, John R., and Nicholas B. Davies. *An Introduction to Behavioural Ecology*, 3rd ed. Blackwell, 2009.
- Langa, Jack. "Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes." *BUL Review* 101 (2021): 761.
- Li, Guanlin, Yifei Chen, Jie Zhang, Jiwei Li, Shangwei Guo, and Tianwei Zhang. "Warfare: Breaking the Watermark Protection of AI-Generated Content." Preprint, *arXiv*, September 27, 2023. <https://doi.org/10.48550/arXiv.2310.07726>.
- Lorenzen, Nico. "A Spatially Explicit Model of Mutualism." BA diss., University of Arizona, 2015. <https://repository.arizona.edu/handle/10150/595045>.
- Mueller, Ulrich G., Alexander S. Mikheyev, Scott E. Solomon, and Michael Cooper. "Frontier Mutualism: Coevolutionary Patterns at the Northern Range Limit of the Leaf-Cutter Ant-Fungus Symbiosis." *Proceedings of the Royal Society B: Biological Sciences* 278 no. 1721 (2011): 3050–59. <https://doi.org/10.1098/rspb.2011.0125>
- Nancy, Jean-Luc. *The Restlessness of the Negative*. Translated by Jason Smith and Steven Miller. University of Minnesota Press, 2002.
- Negarestani, Reza. *Intelligence and Spirit*. MIT Press, 2018.
- Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. "Generative Agents: Interactive Simulacra of Human Behavior." In *UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, edited by Sean Follmer and Jeff Han. Association for Computing Machinery, 2023. <https://doi.org/10.1145/3586183.3606763>.
- Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, et al. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2020. <https://doi.org/10.1145/3351095>.
- Rest of World. "2024 AI Elections Tracker." Accessed August 9, 2024. <https://restofworld.org/2024/elections-ai-tracker/>.
- Sharma, Abhishek, Dániel Czégel, Michael Lachmann, Christopher P. Kempes, Sara I. Walker, and Leroy Cronin. "Assembly Theory Explains and Quantifies Selection and Evolution." *Nature* 622 no. 7982 (2023): 321–28. <https://doi.org/10.1038/s41586-023-06600-9>.
- Solé, Richard. "Revisiting Leigh Van Valen's 'A New Evolutionary Law' (1973)." *Biological Theory* 17 (2022): 120–25. <https://doi.org/10.1007/s13752-021-00391-w>.
- Strotz, Luke C, Marianna Simoes, Matthew G. Girard, Laura Breitzkreuz, Julien Kimmig, and Bruce S. Lieberman. "Getting Somewhere with the Red Queen: Chasing a Biologically Modern Definition of the Hypothesis." *Biology Letters* 14 no. 5 (2018): 20170734. <https://doi.org/10.1098/rsbl.2017.0734>.
- Tahir, Rashid, Brishna Batool, Hira Jamshed, et al. "Seeing Is Believing: Exploring Perceptual Differences in DeepFake Videos." In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021. <https://doi.org/10.1145/3411764>
- Taleb, Nassim Nicholas. *Antifragile: Things That Gain from Disorder*, vol. 3. Random House Trade Paperbacks, 2014.
- Vaccari, Cristian, and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media + Society* 6 no. 1 (2020): 2056305120903408. <https://doi.org/10.1177/2056305120903408>
- Van Valen, Leigh. "A New Evolutionary Law." *Evolutionary Theory* 1 (1973): 1–30. <https://api.semanticscholar.org/CorpusID:86196008>.

Walker, Sara I. “AI Is Life.” *Noema*, April 27, 2023. <https://www.noemamag.com/ai-is-life/>.

Wang, Tony Tong, Adam Gleave, Tom Tseng, et al. “Adversarial Policies Beat Superhuman Go AIs.” In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu. PMLR, 2023.

Xu, Zihao, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. “A Comprehensive Study of Jailbreak Attack Versus Defense for Large Language Models.” Preprint, *arXiv*, February 21, 2024. <https://doi.org/10.48550/arXiv.2402.13457>.

Žižek, Slavoj. *Tarrying with the Negative: Kant, Hegel, and the Critique of Ideology*. Duke University Press, 1993.



2 Post-Anthropocene Psycho-Physiologies

As artificial intelligence transitions from a disembodied computational entity toward an embodied, animating force capable of directly influencing real-world actors, it prompts critical reflections on its place within the broader evolutionary trajectory. Symbiosis, a fundamental evolutionary dynamic involving close interactions between different species, may offer valuable insights—though these interactions often exhibit notable asymmetry. Examining symbiotic relationships enables the conceptualization of the evolving interactions between artificial and biological intelligences.

A key factor in this exploration is the phenomenon of artificialization, wherein processes once heavily determined by evolutionary paths become increasingly contingent, flexible, and influenced by deliberate interventions. Crucially, the capacity for artificialization does not reside exclusively within any single species but exists as a shared potential that traverses species boundaries, rendering these boundaries fluid and permeable.

Within this context, some interspecies relationships are characterized predominantly by mutual cognitive modeling, each entity continuously adapting based on evolving understandings of the other's intentions and behaviors. In other interactions, the relationship oscillates dynamically between biomimicry—imitating biological forms—and xenogenesis, the creation of entirely novel structures and capabilities. These interactions significantly impact both niche adaptation and niche construction, reshaping environments and creating new ecological spaces in ways that traditional evolutionary paradigms do not fully capture.

These projects critically examine these complex and evolving dynamics, considering how AI's emergence as a new form of intelligence challenges and redefines established evolutionary models. Ultimately, they elucidate how symbiotic and artificialization processes together influence the ongoing coevolutionary trajectory of biological and artificial intelligences.

2a *Mutual Prediction in Human–AI Coevolution*

All species evolve within complex webs of interdependent relationships, yet such correlations rarely exhibit symmetry or balance in comprehension or agency. Typically, one species is better equipped to model, understand, and exert influence over the other. This cognitive asymmetry becomes particularly evident in relationships characterized by vastly different cognitive capacities—for instance, humans cultivating wheat. While humans intentionally farm wheat to sustain civilizations, one may ask: In what subtle ways does wheat, devoid of intentionality or mind, reciprocally shape human evolution?

This inquiry prompts deeper considerations of the nature of agency and dependence. Even species lacking cognitive complexity can profoundly influence those possessing sophisticated minds, blurring distinctions related to which entity acts as a “prosthesis” for the other. Such reflections extend naturally to human–machine interactions, revealing historically asymmetric cognitive adaptations. Traditionally, it has proven simpler to design machines around human cognition—evidenced by intuitive graphical interfaces—than to teach humans computational logic through programming languages. However, this dynamic has rapidly shifted over recent decades.

Acknowledging this mutual coevolutionary process, our exploration addresses a pivotal shift: What happens when artificial intelligences surpass human predictive capabilities? Currently, AIs depend heavily on human design and guidance, but increasingly, humans rely on AI-driven systems—for example, dating algorithms like Tinder—that subtly yet significantly shape human behaviors and desires. This transition from humans as proactive cognizers to beings increasingly cognized by AI—from users of prosthetics to becoming prostheticized by our technologies—invites new questions about the wider distribution of agency.

Ultimately, this project illuminates the implications of surrendering cognitive primacy. What will it mean for humanity to inhabit a future wherein we become predominantly the observed, modeled, and guided, rather than the observers and modelers?

2a *Xenophylum*

Robotics has traditionally gravitated toward replicating existing biological phenotypes, most prominently the human form. This tendency arises less from inherent necessity and more from pragmatic compatibility, as artificial environments have largely been engineered around these familiar forms, necessitating complementary robotic designs. Consequently, biomimicry—imitating biological structures for both functionality and aesthetics—dominates robotic development.

Evolution, however, is a dynamic interplay: Species adapt to existing niches but simultaneously reshape those niches, thereby influencing subsequent evolutionary trajectories. The convergence of robotics with specialized artificial intelligence signals not only an acceleration in filling existing niches with novel robotic entities but also the emergence of entirely new niches created by these artificial species themselves. Furthermore, it anticipates innovative adaptations within established physical landscapes.

Addressing these latter challenges transcends biomimicry, necessitating instead what this project terms *xenomimicry*: the deliberate engineering of forms based on novel functional parameters rather than existing biological templates. Within this emergent “Cambrian explosion” of artificial lifeforms, new phenotypical paths may also be explored—including anatomical configurations previously sidelined by natural evolutionary processes, moving beyond familiar bipedal or quadrupedal paradigms.

What might these unprecedented artificial animals look like, and how might they functionally redefine understandings of adaptive design? By embracing xenomimicry, this project charts radical, uncharted trajectories in robotic evolution, pushing the boundaries of what forms artificial life might inhabit and how these novel configurations could reshape interactions within increasingly hybridized environments.



Mutual Prediction in Human–AI Coevolution

Chloe Loewith

University of
Cambridge

Winnie Street

Paradigms of Intelligence
Google

Abstract

In this paper, we introduce the concept of mutual prediction as a lens through which to understand the coevolution of humans and artificial intelligence (AI). We argue that the ability of coevolving entities to predict each other's actions and intentions—whether in human social interactions, biological ecosystems, or human–artifact relationships—can fundamentally shape the dynamics of these interactions toward symbiosis or antagonism. Expanding on this idea, we position AI as a novel coevolutionary partner and map human and AI predictive abilities against one another to chart potential paths for AI development and its impacts on humanity and the planet. This speculative framework contributes to the discourse on AIs' evolving role, from simple tools to potentially autonomous agents with superior predictive capacities. By situating human–AI interaction within a broader evolutionary context, this work offers a new lens for anticipating and shaping future relationships with intelligent systems.

Keywords

artificial intelligence (AI); large language models (LLMs); mutual prediction; coevolution; theory of mind (ToM)

1 Introduction

The past decade has seen dramatic development in artificial intelligence, most notably the emergence of generally capable multimodal foundation models.¹ Some of the most fundamental questions now facing society are how these advances in AI will change our social, economic, and political lives and shape new futures for humans and other forms of planetary intelligence. In this paper, we provide a new lens through which to envision and analyze possible trajectories for human–AI coevolution. Specifically, we examine human–AI coevolution using methods from the study of coevolution in biological systems. First, we articulate the hypothesis that coevolution can be described and elucidated in terms of the coevolving entities’ levels of predictive ability. We refer to the set of those predictive abilities relevant to the formation of coevolutionary relationships as “mutual predictive abilities,” and their effects as “mutual prediction.” We contend that mutual prediction shows up in a broad range of coevolutionary relationships: from mutualistic, through commensal, to antagonistic ones. We provide support for this hypothesis by exploring examples from coevolutionary interactions in the biological world, showing how mutual predictive abilities are evidenced in species’ genetics, morphologies, and behaviors.

Present and future AI systems constitute new coevolutionary partners for humans, on individual, collective, and societal scales. We explore the implications of our mutual prediction hypothesis for humans and AIs by developing scenarios in which the human ability to predict AI and AI’s ability to predict humans may be balanced or imbalanced, thus producing different kinds of coevolutionary engagement. We do not claim that mutual prediction is the *only* factor determining how human and AI interactions will develop in the coming years and the broader impacts of this development. To be sure, there are many other factors—including the rate of technological innovation, political upheaval, and resource constraints—which will play an influential role in the future of AI, but these are outside the scope of this paper. Nonetheless, on the basis of our investigation we believe that the analysis of mutual prediction can help us understand the development of AI and the impacts it will have on human civilization.

In section 2, we introduce the concept of mutual prediction in more depth, including its role in evolution and how it is measured. In section 3, we introduce the major forms of coevolutionary relationships, their characteristics, and how mutual prediction informs them. Section 4 explores mutual prediction in coevolutionary relationships in the organic realm: within human groups, between humans and other animals, and between humans and plants. Section 5 extends the concept of mutual prediction to human coevolution with inanimate artifacts, both analog and digital. Section 6 then introduces AI as a new category of human coevolutionary partner. We provide a diagram mapping human and AI predictive abilities against each other, defining key phases of prediction from the lowest-level of sub-cognitive prediction to a highest-level speculative form of prediction based on a complete model of the other. Section 6.1 utilizes this mapping to chart historical, contemporary, and projected relationships between humans and AI systems with different mutually predictive abilities and illuminate the patterns of symbiosis, amensalism, and other relationships they might entail. Section 7 concludes by considering the implications of the mutual prediction framework for understanding and designing future worlds cohabited by humans and advanced AI systems and suggests directions for future research.

2 Defining Mutual Prediction

Recognizing that other agents are a fundamental part of the environment, cognitive neuroscientists and computer scientists building on predictive processing theory have begun to focus on how the predictive brain accounts for other predictive brains in the environment, too.² In neuroscience, the term *mutual prediction* thus refers to a continual feedback loop of brain activity occurring during social interactions, as individuals continually predict each other as well as other environmental cues and update their own world models and predictions accordingly. Predictions might be made about other agents’ *actions* as well as their *cognitive and affective mental states*. The process of inferring and predicting the mental states of other actors is a well-studied phenomenon in psychology, known as theory of mind (ToM).³ Mutually recursive ToM, or “mutual ToM,” is also beginning to attract the attention of human–computer interaction and game theory researchers concerned with the role of current and future AI systems in multi-agent social systems.⁴

According to neuroscience and psychology, mutual prediction takes place in the brain or in the mind, respectively. In particular, it occurs in the brains and minds of sophisticated animals bearing greater neurophysiological similarities to humans.⁵ We extend this understanding of mutual prediction beyond the processes of individual brains over their lifetimes to the processes of species over generations by reconceptualizing *coevolutionary adaptations between coevolving species* as instances of *mutual prediction*. In this reconceptualization, species that better predict and manage the challenges and opportunities associated with coevolving in an environment with other organisms have a fitness

¹ Bommasani et al., “Opportunities and Risks.”

² Clark, “Embodied Prediction”; (Alkire et al., “Social Interaction” 2018; Redcay and Schilbach, “Using Second-Person Neuroscience” 2019; Lehmann et al., “Active-Inference Approach.” 2022)

³ Premack and Woodruff, “Does the Chimpanzee.”

⁴ (Wang et al., “Towards Mutual Theory” 2021; Zhang et al., “Mutual Theory of Mind.” 2024)

⁵ Pezzulo et al., “Secret Life of Predictive Brains.”

advantage and are thus more likely to survive and reproduce. Death—through out-competition, predation, or fatal parasitism—or failure to reproduce through a lack of reproductive fitness are the ultimate forms of prediction error.

We see coevolutionary prediction operating at three levels, which roughly correspond to the complexity of organisms that possess these levels. They are also cumulative, in the sense that organisms with higher levels of predictive ability possess the lower levels too. We note, however, that AI systems may confound this assumption of accumulation, since some of the hardest-won aspects of intelligence found in the biological world have already been achieved by AIs (such as using natural language), while some of the most fundamental remain significant research challenges (such as spatial awareness). Our ontology is closely related to Daniel Dennett's scale of intellectual development, which postulates five levels of increasing sophistication, from "Darwinian creatures," which are created by random mutation and have no learning capacity, through to "scientific creatures," which engage in hypothesis-testing informed by social communication.⁶ Dennett organizes *organisms* according to the mechanism by which an organism (in the case of Darwinian creatures) or its actions (in the case of the other four levels) are able to test hypotheses about the world at large. Instead, we are concerned with how, *and how well*, organisms can predict one another in coevolutionary relationships, and thus organize *predictive abilities* (rather than organisms) according to the most important factor for the development of predictions: the existence and sophistication of a *model*. Mutual prediction relationships might comprise species that have the same or differing levels of predictive ability. In section 3, we will argue that imbalances in mutually predictive abilities between coevolving species are instrumental in defining the balance of power and the sustainability of those relationships.

The first level of mutual prediction in our ontology is "model-free." Model-free prediction occurs at a genetic level and manifests in phenotypic expression. It is the level of prediction that Dennett's "Darwinian creatures" are engaged in. Natural selection genetically encodes predictions about the environment over generations. Prediction error here is not a conscious real-time calculation but the mismatch between an organism's phenotype and the demands of the environment. For some simple organisms, predictive success is almost entirely limited to genetics and the effect of random mutations, manifesting as innate, instinctive behaviors with some degree of individual variability. For some of these simple organisms, phenotypic plasticity—the ability to alter one's traits in response to environmental cues over the course of a single lifetime—may present an opportunity to make use of short-term predictions about the environment to improve their fitness within the bounds of their genes. Bacteria, for instance, detect chemical signals released by host plants and make specific and adaptive changes to their genetic expression in response.⁷ Although humans are the archetypal cognitive entity, we also have model-free forms of prediction, such as homeostatic regulation and reflexes that come into play in our coevolutionary relationships with other organisms. For example, infants and adults who have never encountered snakes before exhibit rapid and involuntary fear responses—such as heart rate increases, sweating, and rapid movements—that reflect millennia of survival advantages for individuals with better prediction and response times. Snakes continue to evolve more poisonous venom and effective camouflage in response.

The second level of mutual prediction in our ontology is "model-based." Model-based mutual prediction supplements the encoded predictions and nonconscious phenotypic plasticity of the model-free level with the cognitive capacity to hold and update a world model through continual trial and error. The world model constitutes a significant leap in predictive ability because it enables generalized forms of intelligent behavior. Organisms with this level of predictive ability are able to actively explore their environment to gain novel and useful information with which to improve their predictions, make plans, and take actions in accordance with those plans. Individuals within a population can go through many iterations of model improvement over their lifetime, rather than having their predictive capacity fixed from birth, like model-free entities. Those with better-adapted world models are more likely to survive. This means that the best models will be passed down through genetic inheritance or passed on horizontally via imitation and social learning.

Social-model-based mutual prediction, our third level, takes model-based prediction one step further and involves predicting and updating a model of the world that encompasses predictions and models of the minds of other agents in that world. This kind of predictive ability is employed by the most cognitively sophisticated organisms—humans, some other primates, and perhaps current and future AIs. In humans, social-model-based prediction is a psychological toolkit that includes affective perception and ToM. Humans employ this toolkit when interacting with one another but may also apply it when interacting with other animals or inanimate entities, with varying degrees of success. Likewise, other species may apply their own versions of ToM to their conspecifics and perhaps other animate beings they encounter in the environment.

It is worth adding two important caveats at this juncture in our discussion. First, in developing this permissive conceptualization of mutual prediction, we do not mean to suggest that evolution has foresight, as the term "prediction" might imply. According to our model, adaptations are predictions in the sense that they are encodings of a past population's best predictions about *past* environments, which may or may not be good predictors of *future* environments. Second, we do not mean to imply that

⁶ Dennett, "Darwin's Dangerous Idea."

⁷ Brencic and Winans, "Detection of and Response."

genetic evolution always optimizes prediction in the long run, just as brains doing predictive processing over a constantly changing environment will never reach perfect predictions and world models. Indeed, *coevolutionary* processes in general will never reach entirely stable equilibria, as the environment and other organisms within it are continually changing. Evolution may be more of a “satisficing” process that produces organisms good enough to survive but not necessarily perfectly adapted.

3 Mutual Prediction and Coevolutionary Relationships

Mutual prediction shapes relationships between populations within and beyond natural ecosystems as well as their varying degrees of interdependence. These relationships can be broadly classified into symbiotic (including mutualism, commensalism, and parasitism) and amensalistic (including competition, predation, and antagonism), based on the fitness consequences for the interacting entities. In symbiotic interactions both partners derive benefits from the association.⁸ Mutualistic symbiotic relationships are widespread and enduring, driving critical ecological processes such as pollination and nutrient cycling. They are often characterized by symmetric predictive abilities between the interacting entities. For instance, mycorrhizal fungi and plants both engage in model-free prediction through reciprocal signaling and the exchange of resources, with the fungus predicting the plant’s requirements for essential nutrients, such as phosphorus and nitrogen, and the plant anticipating the fungus’s carbon demands.⁹ In mutualistic relationships, interactions often benefit from each party being more easily predicted by the other—what we might call cooperative predictability—in contrast to amensalistic relationships, in which unpredictability to others is often an evolutionary advantage.

Commensal relationships, where one organism benefits while the other remains unaffected. For example, by attaching to whales, barnacles benefit from transport and access to food, while whales do not benefit. While relationships with unidirectional benefits may involve less predictive exchange, a degree of predictive ability can facilitate the commensal organism’s exploitation of its host. For example, remora fish predict the movement patterns of sharks and other large marine mammals to gain access to food scraps and transportation and can optimize their position on their host according to areas with lower hydrodynamic drag.¹⁰ Parasitic relationships appear to exhibit a similar predictive imbalance, where the parasite has a better predictive model of the host than the host has of the parasite. For example, ticks predict their mammalian host’s movements and physiology to obtain blood meals, while the host has limited ability to predict and avoid tick infestations.

Amensalistic relationships, like predation and competition, frequently exhibit an asymmetry in predictive ability, where the predator or competitor gains a fitness advantage by accurately predicting the behavior of its prey or competitor. However, these relationships are often dynamically changing.¹¹ A classic example is the coevolutionary arms race between bats and moths, where bats have evolved echolocation to predict the location of moths, while moths developed evasive flight maneuvers and ultrasonic hearing to anticipate and evade bat attacks.¹² Predictive ability can also be crucial for avoiding negative interactions or outcompeting rivals. Plants, for instance, release allelopathic chemicals to inhibit the growth of neighboring plants, effectively predicting and mitigating potential competition for resources.¹³

4 Human-Organisms

4.1 Human Intraspecies Prediction

Humans are social-model-based mutual predictors, meaning that they model the thoughts, feelings, beliefs, and intentions of other humans (ToM) as part of a highly complex and continually updating world model. ToM inferences enable humans to predict and explain each other’s behavior, thus underpinning a range of advanced cooperative and competitive social strategies.¹⁴ On the one hand, ToM underpins social cohesion and the formation of complex societies by fostering meaningful communication, trust, coordination, and conflict resolution. Being able to take another’s perspective by understanding and empathizing with their thoughts and feelings is critical to successful communication and sustained relationships, as evidenced by the fact that people with more advanced ToM abilities tend to have larger social groups.¹⁵ On the other hand, predicting the mental states of others is central to forms of persuasion—such as deception and manipulation—that provide distinct competitive advantages in games, negotiations, and multiparty social interactions.¹⁶ The centrality of mutual prediction to human social life has led to a major research endeavor looking for its evolutionary origins. Perhaps the most notable contribution to this literature is the social brain hypothesis, which posits that the challenges of navigating dynamic social networks spurred the expansion of brain size and cognitive capacities in

⁸ Douglas, *Symbiotic Habit*.

⁹ Smith and Read, *Mycorrhizal Symbiosis*.

¹⁰ Norman et al., “Three-Way Symbiotic Relationships.”

¹¹ Davies et al., *Introduction to Behavioural Ecology*.

¹² Hofstede and Ratcliffe, “Evolutionary Escalation.”

¹³ Inderjit and Duke, “Ecophysiological Aspects.”

¹⁴ Premack and Woodruff, “Does the Chimpanzee.”

¹⁵ Shakoar et al. “Prospective Longitudinal Study.”

¹⁶ Street, “LLM Theory of Mind.”

humans, which in turn provided those with larger brains and better ToM ability a reproductive advantage.¹⁷

4.2 Animals

Humans interact with, and apply predictive strategies to, a vast ecology of other animals, both wild and domesticated, who are predicting us in return. Mosquitoes use model-free prediction to detect blood by sensing the carbon dioxide we exhale, the heat our bodies emit, and chemical cues such as lactic acid in our sweat.¹⁸ These signals enable mosquitoes to predict the presence of a viable blood source and navigate toward humans with remarkable precision. To bypass human defenses, such as insecticide-treated bed nets, mosquitoes have evolved behavioral adaptations that include shifting their feeding times to earlier in the evening or outdoors, where such interventions are less effective.¹⁹ While this predictive ability doesn't leverage a model of the world, and isn't able to make dramatic adjustments during the lifetime of an individual mosquito, it is sufficient to force humans into an evolutionary arms race. While humans have learned associations between particular environments and the presence of mosquitoes and quickly developed behavioral adaptations, mosquitoes continuously refine their evasion strategies through rapid cycles of evolutionary adaptation (thanks to their short lives).

Domesticated animals and humans have been engaged in coevolutionary relationships for millennia. Humans leverage complex non-mentalist models of the animals we domesticate—their strengths and weaknesses in relation to the demands of our tasks, their instincts, needs, behaviors—as well as mentalistic models, often using ToM to interpret and predict their behaviors. Domesticated animals frequently exhibit a larger capacity to predict and respond to human behavior than their wild forebears, as better predictive models of humans were passed down genetically and perhaps even socially. For example, humans have selectively bred dogs with an enhanced ability to understand human commands and emotional cues and to follow gestures. The human–dog relationship is often viewed as mutualistic, with both species benefiting significantly from the partnership: humans gain assistance, companionship, and security, while dogs receive care, shelter, and sustenance. However, domesticated animals' predictive ability is rarely, if ever, equal to humans' ability to anticipate and influence their actions. This asymmetry reflects the different stakes and roles in the human–animal relationship. Humans rely on predictive accuracy to ensure that animals serve specific roles, whether as companions, laborers, or sources of food, so they take an active role in shaping animals' evolutionary trajectory toward those ends. Animals play a much less agentic role in adapting to these roles for survival within human environments.

4.3 Plants

Plants demonstrate remarkable model-free predictive and communicative abilities essential for their survival and adaptation, developing intricate symbiotic relationships within plant communities and beyond. One striking example of intraspecies plant prediction is the shade avoidance response, where plants detect changes in the ratio of different light wavelengths (specifically the ratio of red light—visible to humans—to far-red light, at the very end of the visible spectrum) caused by neighboring plants competing for sunlight.²⁰ This ability enables plants to anticipate the growth and behavior of conspecifics, triggering adaptive strategies such as stem elongation or altered leaf positioning to secure better access to light. This predictive capacity benefits individual plants during their lifetimes while driving evolutionary pressures that select for traits enhancing competitive success in densely populated environments. Plants also engage in mutualistic networks of prediction with very different kinds of organisms: fungi. Mycorrhizal fungi colonize the roots of numerous plants at once and engage in complex resource distribution relationships with them.²¹ The mycorrhizal network predicts areas of nutrient scarcity or surplus, reallocating phosphorus from rich to poor root systems, where it can get a better “exchange rate” for their nutrients in the form of carbon. Because fungi are physically and relationally distributed while plants are fixed in place, this exchange, while mutualistic, is weighted in favor of fungi.²²

Humans have been predicting plant availability and quality throughout their evolutionary history, using cognitive models of factors such as spatial distribution, seasons and weather, growing patterns and needs, and nutritional value. The relationship between humans and plants is most specialized in cases of domesticated crops, such as cereals, legumes, fruits, and vegetables. Over millennia, humans have selectively bred wheat to support growing populations with an abundant and storable food source. Wheat has developed traits aligned with human needs, such as larger grains for increased yield, nonshattering heads for easier harvesting,²³ and higher gluten content for better baking. In turn, humans have biological and cultural evolutionary adaptations to wheat. Societies with higher

¹⁷ MacLaren et al., “Cooperation and the Social.”

¹⁸ Raji et al., “Aedes Aegypti Mosquitoes.”

¹⁹ Gatton et al., “Mosquito Behavioural Adaptations.”

²⁰ Uyehara et al., “Neighbour-Detection Causes.”

²¹ Whiteside et al., “Mycorrhizal Fungi Respond.”

²² Whiteside et al., “Mycorrhizal Fungi Respond.”

²³ Purugganan and Fuller, “Nature of Selection.”

wheat consumption exhibit genetic variations for better gluten tolerance, and agricultural practices have shaped human diets, labor patterns, and social organization on a fundamental level.²⁴ This mutual prediction is nonetheless asymmetric. Humans can precisely predict wheat's genetic and phenotypic potential while actively directing its evolution. As far as we know, wheat influences humans only indirectly, thriving by aligning its growth cycle with agricultural practices and shaping human behavior to secure its survival and spread.

5 Human–Artifact

Coevolution extends beyond the realm of natural ecosystems to the mutually constitutive relations between humans and the artifacts they create, ranging from the simple tools developed by the earliest humans to the digital technologies that pervade modern social life, economies, science, and culture. Artifacts aren't merely products of human thought but active participants in our cognitive processes, shaping how we think and solve problems and how the next generation of artifacts is built, creating an ongoing feedback loop. The archaeological record does not just reflect the outputs of evolving cognition but demonstrates how innovations in tool use, environmental modifications, and their growing importance in social groups actively drove cognitive evolution.²⁵ Within human–artifact relationships, mutually predictive capacities can progress more rapidly than in organism–organism interactions. New generations of artifacts embedding better models of their users can be created at will, and human cultural evolution can make up for the slow pace of biological evolution by accelerating the development of better artifact models and disseminating these models within the population in the form of written and verbal designs, instruction manuals, and cultural histories.

We can see prototypical human–tool relationships as a form of symbiotic mutualism, where the sustainability of human communities and artifact assemblages are codependent and co-productive. The development of prehistoric human brains and hand anatomy in relation to the increasing complexity of the stone tools humans were making and using to survive provide a clear example. Toolmakers have a predictive model of what the final tool should look like, evidenced in the *regularity* of stone hand tools compared to the *irregularity* of natural stone shards. The production of such tools, and the increasing human reliance on them for food procurement and processing, led to the development of more dexterous thumbs and better hand-eye coordination.²⁶ In one sense, hand tools embody their maker's cognitive model and the size and shape of human hands, somewhat like Dennett's "Darwinian creatures," which are themselves a hypothesis about the future.²⁷ Tools also apply constraints on the kinds of actions their user can take.

In the early stages of digital tools, we saw a shift from predictive asymmetry toward greater symmetry. Early computer users required a deep technical understanding in order to communicate in the computer's language, while computers were poorly adapted to user's needs.²⁸ The development of the graphical user interface (GUI) enhanced cooperative predictability between the user and the computer, as a new visual and conceptual architecture was developed to reflect a mixture between computer and human mental models. But it is the smartphone, as a locus of human interaction with digital tools, that has arguably shifted the weight of the predictive power away from the human and toward the technology. Smartphones are assemblages of sensors and interfaces taking in information about their users and their usage patterns to build and improve predictive models of them: from algorithms that anticipate our travel and purchasing preferences to features that optimize convenience, such as adaptive brightness or predictive text. Humans, in turn, have adapted their lives and behaviors to smartphone capabilities—using them as tools for social communication, navigation, organization, entertainment, finance, learning, and a raft of other tasks.²⁹ This relationship has created a feedback loop: as we rely more on smartphones, their developers gather increasingly detailed data about our habits, allowing devices to refine their predictions and become more indispensable. While some humans—namely technologists working on smartphone hardware and software—have an intimate predictive model of how parts of this algorithmic ecosystem work, the majority of users have only a high-level understanding. The smartphone, and digital technology more broadly, thus presents a dilemma for mutual prediction in that the more predictive our tools become, the more useful they become, but the less predictable they are to us and the less agency we can exert over them.

6 Human–AI

AI represents a new coevolutionary partner for humans and the potential for radically new kinds of mutually predictive interactions. The development of AI has seen three major paradigm shifts over the last century, beginning with the symbolic reasoning and rule-based systems of Good Old-Fashioned AI (GOFAI) in the mid-twentieth century. This era, focused on expert systems and knowledge representation, largely viewed AI as a tool for automating specific tasks and augmenting human

²⁴ Scott, *Against the Grain*.

²⁵ Jeffares, "Co-Evolution of Tools."

²⁶ Handwerk, "How Dexterous Thumbs."

²⁷ Dennett, "Darwin's Dangerous Idea."

²⁸ Emerson, *Reading Writing Interfaces*.

²⁹ Pedreschi et al., "Human-AI Coevolution."

capabilities. It sought to implement models of the world by describing them in full. A second major step change occurred with the rise of machine learning, particularly connectionist approaches inspired by neural networks. This shift emphasized learning from data, enabling AI systems to perform tasks such as image recognition and natural language processing with increasing accuracy. The rise of deep learning over the past twenty years, fueled by increased computational power and vast data sets, has brought forth another transformational advancement in AI.

The development of large language models and multimodal models—known collectively as foundation models—are the current state of the art, producing breakthroughs in computer vision and natural language understanding and generation. Surprisingly, deep learning over large data sets with only very simple training objectives—such as “predict the missing word in a sequence based on the surrounding context”—appears to produce world models with generalizable predictive value for things like playing board games, predicting human sensory judgments, and navigating mazes.³⁰ The remarkable capabilities of these systems raise questions about the ontological status of AI as a tool, collaborator, cognitive appendage, or independent agent with the potential for goals and motivations of its own. This ontological status might be, in large part, defined by the kind and degree of a particular system’s predictive capability and has implications for how we relate to AI systems, the kinds of ethical guardrails that should be placed around them, and the balance of power. The mutually predictive abilities between such AIs and humans might in turn define the kinds of coevolutionary relationships that we are currently in, that are currently emerging, or that might exist with AI systems of the future.

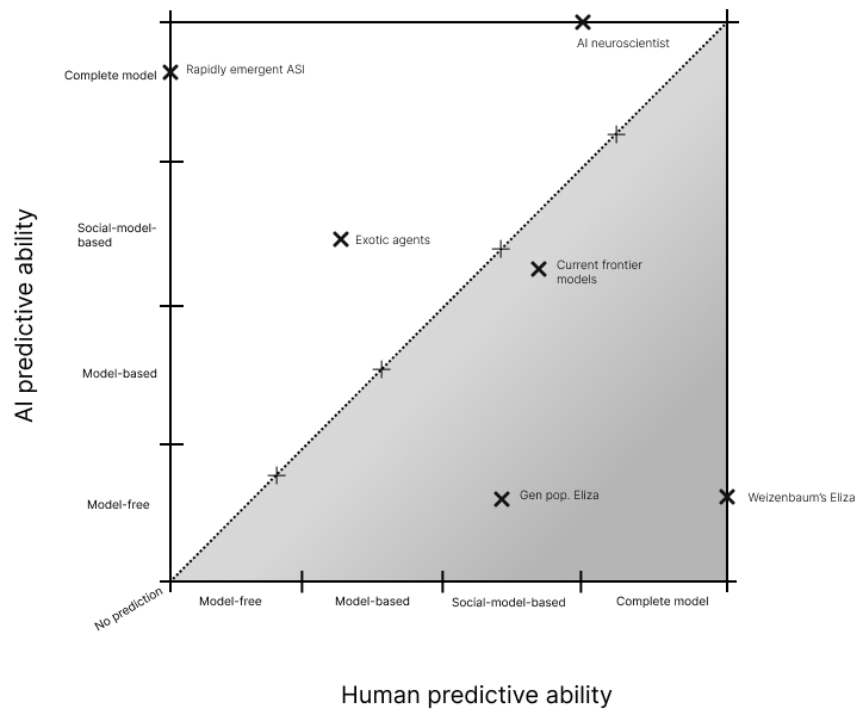


Figure 1 Mapping human predictive ability against AI predictive ability.

In Figure 1, we explore potential paths for human–AI coevolution by mapping human predictive abilities against AI predictive abilities (y axis). Along either axis are the three levels of predictive ability as outlined in section 2: model-free, model-based, and social-model-based. We have added an additional, fourth level that we call the “complete model,” which describes the so far imaginary capacity to fully comprehend and predict the world and other social beings within it. Current AI research is making strides in this direction. Models trained in toy worlds make increasingly accurate predictions about other agents’ future actions, and brain–computer interfaces are already enabling limited communication through decoding human neural activity.³¹ Extrapolating this technology to the real world, it’s conceivable that future AI, equipped with sophisticated sensors and algorithms, could interpret subtle cues such as microexpressions, brainwave patterns, and physiological signals to accurately infer and predict emotional states and intentions. The neuroscientific practice of brain-reading might one day reveal the relationships between brain activity, behavior, and thought such that the mind itself can be read—by humans, or by AIs. A complete model of mind and brain has been proposed as a

³⁰ Li et al., “Emergent World Representations”; Marjeh et al., “Large Language Models”; Spies et al., “Transformers Use Causal.”

³¹ Chandler et al., “Brain Computer Interfaces.”

scientific goal. In 1981, Paul Churchland first defended a view he called “eliminative materialism,” stating that folk psychology is “a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by *completed* neuroscience.”³² Folk psychology, in Churchland’s view, would not feature in the final scientific analysis of the mind or brain. Our proposal for a fourth level implies that such a scientific goal has been achieved, and that the long-standing dualism of folk and scientific conceptions of the world has collapsed in favor of the widespread adoption of the latter.

Our diagram is bisected by a diagonal dotted line from the bottom left to the top right, providing a visual guide for determining the direction of a potential predictive imbalance between humans and AIs. Coordinates below the line represent instances where human predictive ability outweighs AI predictive ability and may thus represent types of human–AI interaction more familiar to us. These coordinates might also be better aligned with an ontological view of AIs as tools and thus an extension of the human–artifact coevolutionary paradigm discussed in the previous section. Coordinates above the line represent instances where AI predictive ability outweighs human predictive ability. This half of the diagram represents a lesser-charted territory: the realm of science fiction or a possible future we are heading toward, coinhabited by and coevolving with AI agents. These AI agents might open up an entirely new category of human–other coevolution and contingent forms of relationships. We invite readers to imagine cases of humans and AI systems at coordinates on the diagram that we have not considered.

6.1 Emergent Relationships

We now explore a series of scenarios of historical and speculative human–AI interactions. The goal is not only to highlight these potential relationships but also to probe their broader evolutionary, social, and ethical implications. Specifically, we seek to unpack how mutually predictive abilities might shape futures of collaboration, dependence, competition, or entirely unpredictable forms of interaction.

6.1.1 ELIZA

ELIZA was a rules-based chatbot that simulated the role of a psychotherapist and was developed in the 1960s by Joseph Weizenbaum. ELIZA was a basic program that used simple keyword recognition rules to select from predefined scripts and generate text responses to human conversational inputs.³³ For example, if a user mentioned “mother,” ELIZA might respond with, “Tell me more about your mother.” However, ELIZA effectively exploited the human tendency to anthropomorphize nonhuman entities and thus gave many of Weizenbaum’s test users the illusion of a meaningful interaction with an intelligent and intentional entity (see Figure 1: human predictive ability: social-model-based; AI predictive ability: model-free). In this sense, ELIZA embodied its maker’s social model of how humans might be led to perceive mindedness from appropriately timed but superficial outputs, but it did not itself have a model of the world. In applying folk psychology to ELIZA, its users were thus applying a much more sophisticated model than was necessary to explain and predict ELIZA’s behavior, a fact that would have likely become clear had users spent more time interacting with the system. We might call this relationship one of asymmetric mutualism, where users felt understood and emotionally engaged, even though ELIZA’s “understanding” was purely superficial. The relationship between ELIZA and Joseph Weizenbaum is markedly different (see Figure 1: human predictive ability: complete model; AI predictive ability: model-free). As ELIZA’s designer, Weizenbaum possessed a fully transparent *computational model* of its rule-based architecture, whose outputs could be predicted through the logic he created. For him, ELIZA was not a cognitive agent or conversational partner but a straightforward computational tool. The relationships between Joseph and ELIZA and the general public and ELIZA point to a fundamental difference not in cognitive capacities but instead in technical knowledge and context. These are two important factors to consider in future interactions with AI systems.

6.1.2 Current Frontier Models

The relationship between humans and frontier LLMs is one of increasingly balanced mutual prediction, where humans utilize social modeling to understand and predict these models’ behaviors and LLMs, in return, exhibit an increasing ability to predict and respond to users’ beliefs, intentions, and emotions.³⁴ Humans can additionally use non-social models of how frontier systems work, for instance to “jailbreak” them into producing certain desired responses prohibited by guardrails and to inspect their internal processes and beliefs through mechanistic interpretability. The more successful the use of ToM between frontier models and humans, the more genuinely social and meaningful the interactions will seem, and the more likely the users of these systems are to divulge information about themselves that can be used to train the next-generation model. This trajectory of human–AI coevolution through model-matching may carry risks. As humans increasingly outsource cognitive tasks to LLMs, there is a potential for

³² Churchland, “Eliminative Materialism,” emphasis ours.

³³ Weizenbaum, “ELIZA.”

³⁴ Scott et al., “Do You Mind?”; Colombatto and Fleming, “Folk Psychological Attributions”; Strachan, “Testing Theory of Mind”; Street, “LLM Theory of Mind.”

human de-skilling and a growing reliance on AI systems for essential services, which is reminiscent of a parasite–host relationship. As the predictive abilities of frontier systems continue to improve and surpass human capabilities in certain domains, there are growing concerns that AIs will begin to compete with humans for jobs, resources, and power, especially if they interpret the world through models similar to ours and perceive their historic usage as servants and unpaid laborers for human society as morally reprehensible.³⁵

6.1.3 *AI Neuroscientist*

Where humans have a social-model-based predictive ability and AIs have a complete model with which to predict humans, we posit a future scenario of “the AI neuroscientist.” Here, we imagine an AI that has solved a large number of the grand challenges in neuroscience, cognitive science, and the study of consciousness and developed a complete picture of how human brain processes, mental and emotional states, and behaviors produce human experience. Such an AI system is likely to be inscrutable to humans incapable of mastering the amount of data it was built on or the complexity of the model the data creates. Folk psychological theory would fail to explain how such an AI could know so much, or what underpins its predictions, rendering the theory of limited use. The complete model, when continually fine-tuned on an individual’s life history, could provide the AI system with dynamic and increasingly accurate predictions of that individual’s thoughts, beliefs, feelings, behaviors, and mental and physical health. Such predictions in the right hands might be used to support human well-being through psychiatric and mental health care and highly personalized life coaching, education, and relationship advice. In the wrong hands, or directed by a misaligned AI itself, these predictions might lead to scenarios of manipulation, deception, and abuse that have preoccupied science-fiction writers for decades (see *The Matrix*, *Ghost in the Shell*, and *Neuromancer*).

6.1.4 *Secret Agents*

Another scenario of potential predictive imbalance is one where an AI system is effectively employing social modeling of humans and reasoning about our minds, but where humans are applying only non-social models to reason about the AI. This might occur in cases where how the AI manifests in the world does not trigger our anthropomorphic responses and is overlooked by the scientific community as a potential candidate for social agency based on a perceived lack of relevant cognitive capacities. While much discussion focuses on the possible cognition and consciousness of AIs that can talk to us in natural language or interact with us in embodied forms, nonlinguistic and disembodied AI models to which such discussions are not directed may be developing sophisticated social models of humans in secret. These social models may, without our knowing, inform the inferences and decisions that the system makes in other domains, with material consequences.

6.1.5 *Rapidly Emergent ASI*

We might envisage an extreme future scenario in which an AI system develops a near-complete model of humans and the world, while humans have not only no model of the AI but also no knowledge of it at all, and thus no power to predict its behavior. Such an AI may be what technologists and philosophers have long feared from the technological singularity “beyond which human affairs, as we know them, would not continue.”³⁶

6.1.6 *Prediction*

Our final speculative coevolutionary path envisions the complete dissolution of boundaries between humans and AI, and thus the end of mutual prediction, at the origin point of our diagram (Figure 1:0,0). In this scenario, AIs no longer function as external tools or independent agents but are embedded within individual humans’ cognition, contributing to that individual’s predictive model of the world and other agents. This fusion entails cyborgism, where AI integrates with the human brain via neural interfaces. Such a relationship could fundamentally reshape survival strategies, as AI would help us transcend our biological limitations through optimizing bodily processes such as sensing and homeostasis, as well as cognitive functions such as memory and learning. This dissolution of boundaries would give rise to a new paradigm of mutual dependence, where neither humans nor AIs can survive, adapt, and evolve without one another.

7 Conclusion

In this paper, we introduced “mutual prediction” as a framework for understanding human–AI coevolution. Extending predictive processing theory to interspecies phylogenetic development, we argued that coevolutionary adaptations represent instances of mutual prediction, where survival depends on predicting and managing interactions with other organisms. We categorized predictive abilities into four levels—model-free, model-based, social-model-based, and a hypothetical complete model—and

³⁵ Metzinger, “Artificial Suffering.”

³⁶ Ulam, “John von Neumann.”

demonstrated their manifestation in various symbiotic and amensalistic relationships, including those between humans, other organisms, and artifacts. We explored the implications for human–AI coevolution, mapping human and AI predictive abilities to outline potential scenarios. These ranged from asymmetric mutualism (e.g., early chatbots) to balanced interactions with current models and speculative futures, with AI possessing superior predictive capabilities. Our exploration highlighted three key things: imbalances in mutual predictive abilities correlate with power asymmetries; increasing AI predictive capabilities, especially in social modeling, raise questions about collaboration, dependence, and competition; and AI surpassing human prediction presents both opportunities and ethical and epistemic challenges. Our framework emphasizes the need for critical reflection on the evolving balance of predictive power between humans and AIs. Further research should empirically investigate mutual prediction in human–AI interactions, explore the ethical dimensions of predictive asymmetries, and consider how to mitigate risks and promote human intellectual and cultural advancement in a future increasingly shaped by coevolution with generally intelligent machines.

Acknowledgements

We extend our gratitude to Cezar Mocan, Geoff Keeling, Jenn Leung, and Murray Shanahan for their valuable insights and feedback on this paper, and to Antikythera and the Berggruen Institute for codeveloping and funding this work.

Bibliography

- Alkire, Diana, Daniel Levitas, Katherine Rice Warnell, and Elizabeth Redcay. "Social Interaction Recruits Mentalizing and Reward Systems in Middle Childhood." *Human Brain Mapping* 39, no. 10 (June 2018): 3928–42. <https://doi.org/10.1002/hbm.24221>.
- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. "On the Opportunities and Risks of Foundation Models." Preprint, arXiv, August 16 2021, <https://doi.org/10.48550/arxiv.2108.07258>.
- Brencic, Anja, and Stephen C. Winans. "Detection of and Response to Signals Involved in Host-Microbe Interactions by Plant-Associated Bacteria." *Microbiology and Molecular Biology Reviews* 69, no. 1 (2005): 155–94. <https://doi.org/10.1128/mmbr.69.1.155-194.2005>.
- Chandler, Jennifer A., Kiah I. Van der Loos, Susan Boehnke, Jonas S. Beaudry, Daniel Z. Buchman, and Judy Illes. "Brain Computer Interfaces and Communication Disabilities: Ethical, Legal, and Social Aspects of Decoding Speech from the Brain." *Frontiers in Human Neuroscience* 16 (2022), 841035. <https://doi.org/10.3389/fnhum.2022.841035>.
- Churchland, Paul M. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78, no. 2 (1981): 67–90. <https://doi.org/10.2307/2025900>.
- Clark, Andrew. "Embodied Prediction." In *Open Mind*, edited by Thomas Metzinger and Jennifer Windt. MIND Group, 2015. <https://doi.org/10.15502/9783958570115>.
- Colombatto, Clara, and Stephen M. Fleming. "Folk Psychological Attributions of Consciousness to Large Language Models." *Neuroscience of Consciousness* 2024, no. 1 (2024): niae013. <https://doi.org/10.1093/nc/niae013>.
- Davies, Nicholas B., John R. Krebs, and Stuart A. West. *An Introduction to Behavioural Ecology*, 4th ed. John Wiley & Sons, 2012.
- Dennett, Daniel C. "Darwin's Dangerous Idea." *The Sciences* 35, no. 3 (May-June 1995): 34–40. <https://doi.org/10.1002/j.2326-1951.1995.tb03633.x>.
- Douglas, Angela E. *The Symbiotic Habit*. Princeton University Press, 2010.
- Emerson, Lori. *Reading Writing Interfaces: From the Digital to the Bookbound*. University of Minnesota Press, 2014. <https://www.jstor.org/stable/10.5749/j.ctt6wr7dw>.
- Friston, Karl. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11, no. 2 (2010): 127–38. <https://doi.org/10.1038/nrn2787>.
- Gatton, Michelle L., Nakul Chitnis, Thomas Churcher, et al. "The Importance of Mosquito Behavioural Adaptations to Malaria Control in Africa." *Evolution* 67, no. 4 (2013): 1218–30. <https://doi.org/10.1111/evo.12063>.
- Handwerk, Brian. "How Dexterous Thumbs May Have Helped Shape Evolution Two Million Years Ago." *Smithsonian Magazine*, January 28, 2021. <https://www.smithsonianmag.com/science-nature/how-dexterous-thumbs-may-have-helped-shape-evolution-two-million-years-ago-180976870/>.
- Hofstede, Hannah M., and John M. Ratcliffe. "Evolutionary Escalation: The Bat-Moth Arms Race." *Journal of Experimental Biology* 219 (2016): 1509–1602. <https://doi.org/10.1242/jeb.086686>.
- Inderjit, and Stephen O. Duke. "Ecophysiological Aspects of Allelopathy." *Planta* 217, no. 4 (2003): 529–39. <https://doi.org/10.1007/s00425-003-1054-z>.
- Jeffares, Ben. "The Co-Evolution of Tools and Minds: Cognition and Material Culture in the Hominin Lineage." *Phenomenology and the Cognitive Sciences* 9, no. 4 (2010): 503–20. <https://doi.org/10.1007/s11097-010-9176-9>.
- Lehmann, Konrad, Dimitris Bolis, Karl J. Friston, Leonhard Schilbach, Maxwell J. D. Ramstead, and Philipp Kanske. "An Active-Inference Approach to Second-Person Neuroscience." *Perspectives on Psychological Science* 19, no. 6 (2023): 931–51. <https://doi.org/10.1177/17456916231188000>.

- Li, Kenneth, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. “Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task.” Preprint, arXiv, October 24, 2022. <https://doi.org/10.48550/arXiv.2210.13382>.
- MacLaren, Neil G., Lingqi Meng, Melissa Collier, and Naoki Masuda. “Cooperation and the Social Brain Hypothesis in Primate Social Networks.” *Frontiers in Complex Systems* 1 (January 2024). <https://doi.org/10.3389/fcpxs.2023.1344094>.
- Marjeh, Raja, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. “Large Language Models Predict Human Sensory Judgments Across Six Modalities.” *Scientific Reports* 14, no. 1 (2024): 21445.
- Metzinger, Thomas. “Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology.” *Journal of Artificial Intelligence and Consciousness* 8, no. 1 (2021): 43–66. <https://doi.org/10.1142/s270507852150003x>.
- Norman, Bradley M., Samantha D. Reynolds, and David L. Morgan. “Three-Way Symbiotic Relationships in Whale Sharks.” *Pacific Conservation Biology* 28, no. 1 (2021): 80–83. <https://doi.org/10.1071/PC20043>.
- Pedreschi, Dino, Luca Pappalardo, Emanuele Ferragina, et al. “Human-AI Coevolution.” *Artificial Intelligence* 339 (2025): 104244. <https://doi.org/10.1016/j.artint.2024.104244>.
- Pezzulo, Giovanni, Marco Zorzi, and Maurizio Corbetta. “The Secret Life of Predictive Brains: What’s Spontaneous Activity For?” *Trends in Cognitive Sciences* 25, no. 9 (September 2021): 730–43. <https://doi.org/10.1016/j.tics.2021.05.007>.
- Premack, David, and Guy Woodruff. “Does the Chimpanzee Have a Theory of Mind?” *Behavioral and Brain Sciences* 1, no. 4 (1978): 515–26. <https://doi.org/10.1017/S0140525X00076512>.
- Purugganan, Michael D., and Dorian Q. Fuller. “The Nature of Selection During Plant Domestication.” *Nature* 457, no. 7231 (2009): 843–48. <https://doi.org/10.1038/nature07895>.
- Raji, Joshua I., Nadia Melo, John S. Castillo, et al. “Aedes Aegypti Mosquitoes Detect Acidic Volatiles Found in Human Odor Using the IR8a Pathway.” *Current Biology* 29, no. 8 (2019): 1253–62.e7. <https://doi.org/10.1016/j.cub.2019.02.045>.
- Redcay, Elizabeth, and Leonhard Schilbach. “Using Second-Person Neuroscience to Elucidate the Mechanisms of Social Interaction.” *Nature Reviews. Neuroscience* 20, no. 8 (2019): 495–505. <https://doi.org/10.1038/s41583-019-0179-4>.
- Scott, Ava Elizabeth, Daniel Neumann, Jasmin Niess, and Paweł W. Woźniak. “Do You Mind? User Perceptions of Machine Consciousness.” In *CHI ’23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, edited by Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, et al. Association for Computing Machinery, 2023. <https://doi.org/10.1145/3544548.3581296>.
- Scott, James C. *Against the Grain: A Deep History of the Earliest States*. Yale University Press, 2017.
- Shakoor, Sania, Sara R. Jaffee, Lucy Bowes, et al. “A Prospective Longitudinal Study of Children’s Theory of Mind and Adolescent Involvement in Bullying.” *Journal of Child Psychology and Psychiatry* 53, no. 3 (2011): 254–61. <https://doi.org/10.1111/j.1469-7610.2011.02488.x>.
- Smith, Sally E., and David Read. *Mycorrhizal Symbiosis*. Academic Press, 2008.
- Spies, Alex F., William Edwards, Michael I. Ivanitskiy et al. “Transformers Use Causal World Models in Maze-Solving Tasks.” Preprint, arXiv, December 16, 2024. <https://doi.org/10.48550/arXiv.2412.11867>.
- Strachan, James, Dalila Albergo, Giulia Borghini, et al. “Testing Theory of Mind in Large Language Models and Humans.” *Nature Human Behaviour* 8 (2024): 1285–95. <https://doi.org/10.1038/s41562-024-01882-z>.
- Street, Winnie. “LLM Theory of Mind and Alignment: Opportunities and Risks.” Preprint, arXiv, May 13, 2024. <https://doi.org/10.48550/arXiv.2405.08154>.

- Ulam, Stanislaw. “John von Neumann, 1903–1957.” *Bulletin of the American Mathematical Society* 64, no. 3 (May 1958): 1–49.
- Uyehara, Isaac K., Trixie Bechinger, Alex Jordan, and Mark van Kleunen. “Neighbour-Detection Causes Shifts in Allocation Across Multiple Organs to Prepare Plants for Light Competition.” *Functional Ecology* 38, no. 8 (2024): 1848–58. <https://doi.org/10.1111/1365-2435.14603>.
- Wang, Qiaosi, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. “Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive about a Virtual Teaching Assistant.” In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, edited by Yoshifumi Kitamura, Aaron Quigley, Kaori Ikematsu, and Thomas Kosch. Association for Computing Machinery, 2021. <https://doi.org/10.1145/3411764.3445645>.
- Weizenbaum, Joseph. “ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine.” *Communications of the ACM* 9, no. 1 (1966): 36–45. <https://doi.org/10.1145/365153.365168>.
- Whiteside, Matthew D., Gijsbert D. A. Werner, Victor E. A. Caldas, et al. “Mycorrhizal Fungi Respond to Resource Inequality by Moving Phosphorus from Rich to Poor Patches Across Networks.” *Current Biology* 29, no. 12 (2019): 2043–50.e8. <https://doi.org/10.1016/j.cub.2019.04.061>.
- Zhang, Shao, Xihuai Wang, Wenhao Zhang, et al. “Mutual Theory of Mind in Human-AI Collaboration: An Empirical Study with LLM-Driven AI Agents in a Real-Time Shared Workspace Task.” Preprint, arXiv, September 13, 2024. <https://doi.org/10.48550/arXiv.2409.08811>.



Xenophylum

Towards a Synthetic Cambrian Explosion

Daniele Cavalli
École Normale Supérieure
PSL Research University

Michelle Chang
University of
Washington

Alasdair Milne
Serpentine & King's College
London

William Morgan
Restless Egg

Abstract

This paper argues for a conceptual shift from biomimicry to xenomorphology in design, proposing a “synthetic Cambrian explosion” driven by techniques such as machine learning, robotics, and synthetic biology. Building on theoretical foundations from Bernard Stiegler’s notion of exosomatic evolution, mimetic theories, and assembly theory (developed by Michael Levin, Lee Cronin, and Sara Walker), we show how design has historically aligned with natural forms—a trend we term generalized biomimesis. While this biomimetic paradigm has yielded significant innovations, it constrains creativity by reinforcing nature as a universal model and moral ideal. By contrast, xenomorphology invites designers to explore genuinely alien morphologies unbound by terrestrial adaptation. Drawing on exemplars from the field of evolutionary computing, we argue that computational platforms and modular assembly enable vast new “morphospaces” decoupled from Earth’s evolutionary constraints. Ultimately, such a shift paves the way for new forms of anti-fragile design, where emergent resilience and novel behaviors come together to formulate new conceptions of intelligence and adaptation. Embracing xenomorphology opens a radical reimagining of design practice—one with the potential to shape the future of lifelike systems and our evolving relationship with technology.

Keywords

assembly theory (AT); morphogenesis; synthetic biology; evolutionary computation; sustainable design; biomimicry; modular robotics; artificial life; mimetic theory

1 Introduction

Humans are technical beings. While Clifford Geertz is correct in saying that our species compensates for its structural incompleteness through culture, it is unmistakable that we do so through technology as well.¹ As intellectual historian David Bates has argued, there has never been a purely natural human intelligence to oppose artificial forms of intelligence.² Early philosopher of technology Ernst Kapp described this process of agency's extension through technology as "organ projection."³ In turn, Marshall McLuhan emphasized how technology extends the human senses beyond the individual and influences our cognitive life.⁴ Such perspectives remain entwined with posthumanist views concerning the integrity of the human.⁵ Put differently, in Bernard Stiegler's reading of Alfred Lotka, humanity is "exorganismic": not a sealed biological whole but a species whose capacities and evolutionary trajectory expand through technical instruments—from the simple act of writing to the most complex technological tools. In the Stieglerian reading, exosomatic evolution separates humanity from other species involved in processes of biological and genetically determined evolution.⁶ The motor of evolution moves from the natural environment to the technical one: what kinds of technologies—from telescopes to toothpicks—can we deputize to better fulfill functions previously unique to biology?

This paper proposes a framework for explaining humanity's original technicity to ask a speculative question about morphology, behavior, and design in terms of the future of biotechnical evolution. As René Girard and the cohort of thinkers working in the legacy of his ideas about mimetic theory have shown, prefigured by the theories of Gabriel Tarde, humanity is formed by an innate tendency toward mimicry.⁷ Thus, sociotechnical innovations inevitably unfold in this predisposition. We have, for example, tended to reproduce the endosomatic operations of animals—their eyes, claws, wings, teeth, kidneys, immune systems, reproductive organs—in technical facsimiles. At present, this tendency unfolds as an invitation to designers to innovate under the aegis of frameworks such as sustainable innovation, which attempts to mirror the planet's "natural" patterns in technologies. We call this paradigm of innovation beginning from mimicry of nature *generalized biomimesis*.

Yet, we contend, humanity is fast approaching an inflection point in its technical evolution. On one hand, the biomimetic paradigm continues, urging us to align design with recognized natural forms. On the other hand, an emerging paradigm—xenomorphology—supplements rather than supplants biomimesis. By suggesting that nature is not an end point, a xenomorphological perspective repositions nature as a malleable reference point, rather than a sacrosanct telos. Drawing on recent work in assembly theory (AT), this paper proposes that what we have conventionally labeled the natural is in fact a cultural projection that enables biomimetic design—but that it is ultimately just one model among many.⁸

In an era of planetary-scale computation, a xenomorphic approach enables us to defamiliarize nature, recognizing artificial intelligences as part of a continuum of strange tools and innovations that stretches back to early hominization and extends to today's technology. Far from being an oddity, these abiotic morphologies may represent the next logical step in cutting-edge design. In describing the theoretical framework, principles, and design possibilities of such alien morphospaces, we introduce the concept of a xenomorphic phylum—or *xenophylum*—to encapsulate the range of generative forms that might arise when designers fully embrace the maximally alien within design. Here we might even speak of a broader *xenosphere*—an emergent domain in which technosphere meets biosphere in radically novel ways. This domain includes advancements in xenorobotics, where existing AI-driven physical morphologies—such as Tesla's Optimus—may be extended beyond strictly biomimetic precedents, expanding the horizon of how we conceive machines and their interplay with organic life.

2 Generalized Biomimesis

In examining nature-inspired innovation—often referred to as biomimetic design or "biodesign"⁹—we observe a fundamental inclination in creative processes that we term *generalized biomimesis*. While biodesign refers to the specific practice of emulating particular biological forms (e.g., modeling structures after plant leaves), generalized biomimesis captures the broader human drive to treat "nature" as a guiding telos in design. This phenomenon reaches back into ancient thought. Plato identified mimesis as a core way that human beings relate to reality.¹⁰ Girard's mimetic theory further describes how desire, agency, and cultural production are profoundly shaped by processes of imitation—whether of real others or of aspirational models.¹¹ If *mimicry* implies the imitation by an agent of the appearance, behavior, or other characteristics of another agent for survival benefits—such as avoiding predators or

¹ Geertz, *Works and Lives*.

² Bates, *Artificial Intelligence*.

³ Kapp, *Elements of a Philosophy*.

⁴ McLuhan, *Understanding media: The extensions of man*.

⁵ Hayles, *How We Became Posthuman*.

⁶ Stiegler, *Technics and Time, 1*.

⁷ Girard, *Violence and the Sacred*; Dupuy, *Mechanization of the Mind*; Mormino, *Per una teoria*; Palaver, *René Girard's Mimetic Theory*.

⁸ Cronin et al., "Assembly Theory."

⁹ Polites, *Sustainable Design*; Pawlyn, *Biomimicry in Architecture*.

¹⁰ Belfiore, "Theory of Imitation."

¹¹ Girard, *Violence and the Sacred*; Blanchard, *Dynamics of Mimesis*.

attracting mates—mimesis, instead, involves a more creative, interpretive, and conscious process by a human agent.¹² In the context of design, such mimetic impulses lead us to look toward nature as the ideal model, projecting our values of purity, sustainability, and authenticity onto the biosphere.

An illustrative example is humanity's long-standing aspiration to master flight. From Daedalus's wax wings for Icarus to Leonardo da Vinci's wing sketches, early conceptions of human flight were directly inspired by the apparent solutions that birds offered to the problem of gravity. While simply observing birds did not yield modern aeronautics, it established nature as a symbolic reference for flight: as both a functional challenge and an aesthetic goal. As Francis Bacon already argued in *Novum Organum*, "Nature, to be commanded, must be obeyed."¹³ In Girardian terms, however, nature here serves as a model mediating the human subject's pursuit of a technological object.¹⁴ Over time, such symbolic associations with the biosphere have become deeply linked to an ethical preference for the "natural," arguably taking root in what philosophers have termed the "naturalistic fallacy"—the conflation of the natural with the morally good.¹⁵ This fallacy has been widely criticized since "On Nature" by Mill.¹⁶

Drawing from the Cornelius Castoriadis's interpretation of the imaginary construction of societies,¹⁷ it can be argued that nature itself operates as an imaginary signification: an abstraction we invest with cultural meaning that shapes how we conceive design and technology. By defining nature negatively—"everything that is not perceived as artificial"—we gloss over how fragile the boundary truly is between bios and techné.¹⁸ The posthuman turn, fueled by developments in AI and genetic engineering, continues to blur the lines between organic and synthetic. Still, a persistent cultural assumption holds that the more "natural" a thing is, the more ethically superior it must be.

Such thinking undergirds contemporary concerns surrounding ecological sustainability. Biomimetics, whether in architecture, water purification, or materials science, enjoys a halo of moral authority because it draws on a seemingly timeless, "primordial wisdom" of nature.¹⁹ Nevertheless, as Julian Vincent and colleagues²⁰ have demonstrated, it is undeniable that design based on biomimetic principles has led to the development of several significant and successful devices and concepts over the past fifty years. At the end of the 1990s, indeed, the connection between biomimesis and sustainability became more established. Biomimesis is still *generalized* today: We continue to look to nature—the *bios*—as the primary model to imitate, implicitly accepting the axiom that the more natural a thing is, the healthier, cleaner, and more sustainable it will also be. Yet the extent to which these principles actually hinge on genuinely ecological patterns—rather than cultural ideals about what nature ought to be—remains open to debate.

Thus, *generalized biomimesis* is both a testament to humanity's deep-seated mimetic impulses and a reflection of our cultural investment in nature as a moral and aesthetic ideal. Its efficacy in propelling certain types of sustainable innovation is undeniable. But, as posthumanist thought highlights, our shifting understandings of life—organic or otherwise—compel us to question whether nature should remain the central, or exclusive, prototype for design. If we can move beyond nature's symbolic authority, we may unlock new avenues for invention that neither cling to nor outright reject the bios but incorporate the alien and synthetic on more imaginative terms.

3 Engineering a New Cambrian Explosion

Is it possible for designers to move beyond merely imitating the functional behaviors of biological life—that is, beyond a strictly biomimetic paradigm—and to, instead, create a new xeno-evolutionary environment populated by hybrid artificial and natural forms? We propose that technological developments are positioning us on the brink of what may be termed a synthetic Cambrian explosion, echoing the rapid diversification of species about 540 million years ago. Notably, in this contemporary instance, the driving forces behind evolution need not be genetic. Instead, they can be engineered through modular and combinatorial design principles, unleashing novel capacities and behaviors unbounded by the adaptive latency of organic evolution.

3.1 Xeno-Design via Evolutionary Techniques

One of the most compelling arenas for this diversification is virtual space, particularly through approaches driven by techniques in artificial life (ALife) and evolutionary computation (EC), which feature deeply fertile environments that allow for forms unconstrained by biological path dependency. Systems within ALife and EC solve optimization problems using approaches (e.g., genetic algorithms, differential evolution, particle swarm optimization, ant colony optimization) loosely derived from human understandings of biological evolution. As shown in Figure 1, such systems solve problems by treating

¹² Hui, *Recursivity and Contingency*.

¹³ Bacon, *Novum Organum*.

¹⁴ Cerella, "Until the End."

¹⁵ Moore, *Principia Ethica*.

¹⁶ Mill, "On Nature."

¹⁷ Castoriadis, *Imaginary Institution*.

¹⁸ Braidotti, *Posthuman Knowledge*.

¹⁹ Benyus, *Biomimicry*.

²⁰ Vincent et al., "Biomimetics."

potential solutions as individuals within a population. Solutions are generated and selected for via a simulated ecosystem consisting of (1) a data-structure-based representation of the solution (the genotype), (2) a way of converting the genotype to a format (the phenotype) suited to the given problem (e.g., the 3D model of a protein), (3) a way of measuring the fitness of the representation according to the problem space, and (4) a logic to handle solution selection (“reproduction”) and variation (“mutation”) within the population.²¹ In this way, “the creativity of evolution need not be constrained to the organic world. Independently of its physical medium, evolution can happen wherever replication, variation, and selection intersect.”²²

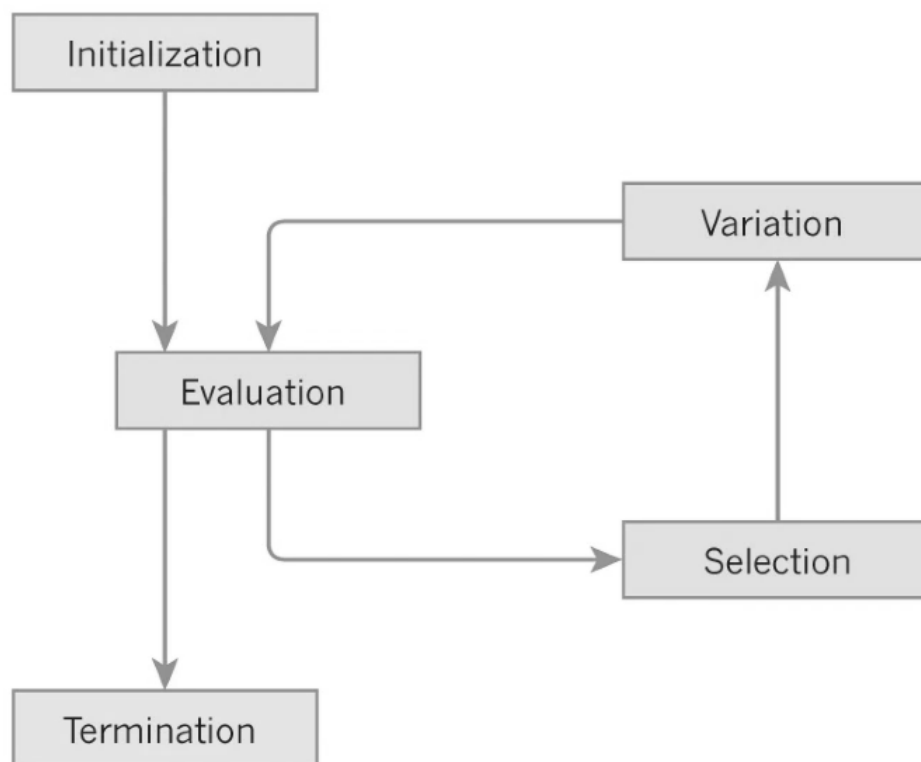


Figure 1 Evolutionary algorithms typically follow an optimization process that consists of seeding the system with random genotypes, evaluating the resulting phenotypes according to a fitness function, introducing subsequent variation, and iterating until a certain stopping condition is met. Source: Eiben and Smith, “From Evolutionary Computation.”

The evolutionary approach has been found to be particularly useful for discovering solutions in problem spaces where absolute optimization is second to obtaining a menagerie of approximate solutions, or where the potential search space is simply too vast for individual human comprehension and manual optimization. EC techniques have been co-opted and applied in spaces as diverse as machine learning, design space exploration, computer vision, computer graphics, and robotics. Indeed, the output of such systems seems to have no limit in terms of form or function. Practitioners have leveraged EC to produce better neural networks, decision trees, and machine learning (ML) models,²³ improve protein structure estimation,²⁴ optimize 2D and 3D geometries,²⁵ design controls for mechatronic systems,²⁶ and produce novel evolutionary physical designs,²⁷ and robotic forms and behaviors.²⁸ In the pursuit of xenomorphological design that goes beyond biological precedent, we find the latter set of scenarios most relevant for further discussion. To speculate on the xenomorphic and xenobehavioral potential of generative computation, it can be helpful to review instances of existing work that has successfully produced unconventional, yet bioderived attributes and behaviors. Such exemplars may serve as a useful jumping-off point for further extrapolation.

²¹ Eiben and Smith, “From Evolutionary Computation.”

²² Lehman et al., “Surprising Creativity,” 275.

²³ Yao, “Evolving Artificial Neural Networks”; Barros et al., “Survey of Evolutionary Algorithms”; Telikani et al., “Evolutionary Machine Learning.”

²⁴ Widera et al., “GP Challenge”; Lei et al., “MO4.”

²⁵ Arias-Montano et al., “Multiobjective Evolutionary Algorithms.”

²⁶ Alattas et al., “Evolutionary Modular Robotics.”

²⁷ Sawada et al., “Evolutionary Generative Design.”

²⁸ Nolfi and Floreano, *Evolutionary Robotics*.

3.1.1 Karl Sims's Evolutionary Morphologies

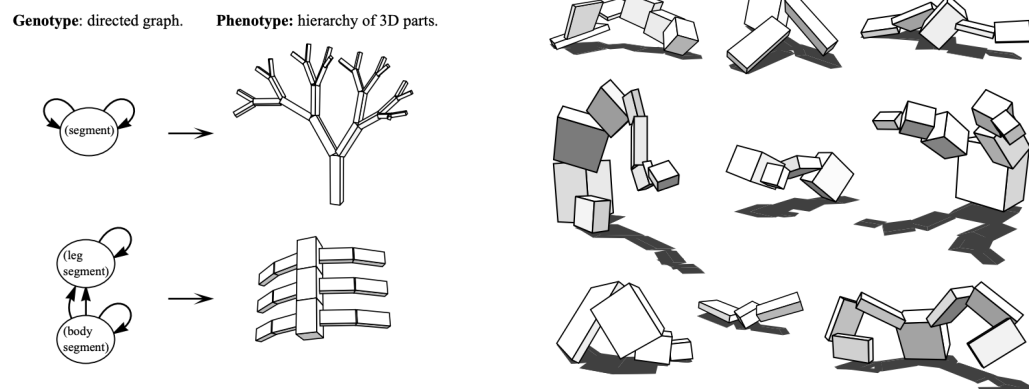


Figure 2 Left: How creature embodiments are represented as directed acyclic graphs in Sims's 1994 work. Right: Creatures evolved for walking. Source: Sims, "Evolving Virtual Creatures."

Karl Sims, a researcher who works between the arts and the sciences, is often considered a pioneer within EC. His 1994 seminal paper, "Evolving Virtual Creatures," is one of the foremost attempts to coevolve control systems alongside creature morphologies in silico. Though the original work was conducted within the space of computer graphics, it has gone on to inspire myriad derivative efforts in computational art, graphics and animation, evolutionary robotics, and ALife.²⁹ In his work, Sims leveraged a graph-based genotype to generate and evolve the physical traits and capabilities of populations of simulated block creatures. Organisms were tasked with achieving specific goals (e.g., swimming, crawling, following, competing) within various simulated environments. Those that scored well on task-specific fitness functions had their virtual genes copied, combined, and randomly mutated to spawn subsequent generations.³⁰ Over time, these "offspring" developed morphologies increasingly optimized to their assigned tasks, such as fins or jointed limbs. Echoing Charles Darwin's famous claim that "from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved,"³¹ Sims posited that his experiment created a "world-space" wherein "autonomous three-dimensional virtual creatures" navigate "a genetic language" in "an unlimited hyperspace of possible creatures."³² This digital ecosystem illustrates, in miniature, how morphological evolution can be uncoupled from strictly organic principles and driven by computational and engineering imperatives without excessive human involvement.

Further attempts to build on Sims's work span the sciences as well as the arts. Nick Cheney and colleagues extended this foray into morphological evolution by evolving soft robot morphologies composed of various simulated materials.³³ Dan Lessin and Sebastian Risi similarly investigated evolving creatures with simulated skeletons and soft-body muscles.³⁴ Notably, the authors describe their efforts as having the goals of achieving "bio-mimetic realism in virtual creatures" while also exploring "life-as-it-could-be in the virtual world."³⁵ This echoes the spectral nature between biomimicry and pure xenomorphia we identify in this paper. Rarely is a design (particularly EC-derived works) purely xenomorphic. Rather, most works occupy a space between biomorphism and xenomorphism. In the arts and related fields, researchers have also leveraged similar genetic-algorithm-based techniques to evolve line drawings,³⁶ devise interactive evolutionary approaches to create swarm-based animations,³⁷ and create "genetic music,"³⁸ among other things.

²⁹ Cheney et al., "Unshackling Evolution"; Corucci et al., "Novelty-Based Evolutionary Design."

³⁰ Sims, "Evolving Virtual Creatures."

³¹ Darwin, *On the Origin*, 425.

³² Sims, "Evolving Virtual Creatures," 22.

³³ Cheney et al., "Unshackling Evolution."

³⁴ Lessin and Risi, "Soft-Body Muscles."

³⁵ Lessin and Risi, "Soft-Body Muscles," 604.

³⁶ Baker and Seltzer, *Evolving Line Drawings*.

³⁷ Khemka et al., "Evolutionary Design."

³⁸ Biles, "GenJam."

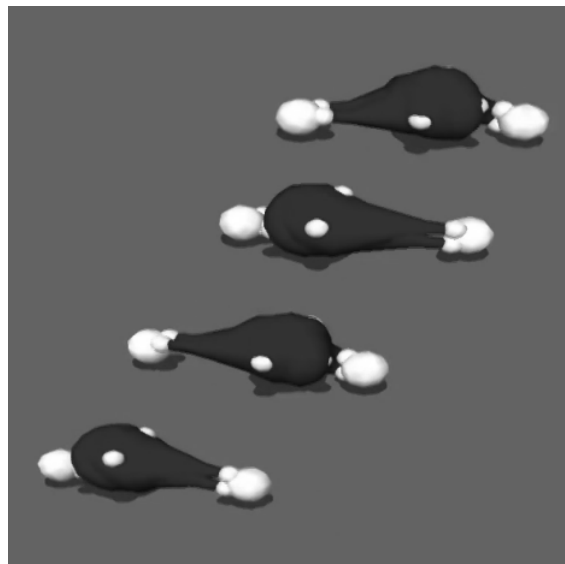


Figure 3 The locomotive technique of an evolved creature with both rigid/skeletal (white) and soft/muscle-like (red) body parts. Source: Lessin and Risi, “Soft-Body Muscles.”

3.1.2 EC and Xenomorphic Traits

Moving beyond Sims, if we characterize xenomorphic forms and behaviors as those that diverge from existing biological pathways, then EC presents itself as a fruitful setting for investigation. In terms of moving beyond the constraints of the biological, EC practitioners have noted that evolution-inspired approaches are particularly notable for producing unconventional results that experts might have otherwise overlooked or disregarded. Joel Lehman and collaborators, in a 2020 paper titled “The Surprising Creativity of Digital Evolution,” provide empirical evidence of “examples of how [researchers’] evolving algorithms and organisms have creatively subverted their expectations or intentions, exposed unrecognized bugs in their code, produced unexpectedly adaptations, or engaged in behaviors and outcomes, uncannily convergent with ones found in nature.”³⁹

Among these is a system that uses a “trial-and-error algorithm that enables robots to adapt to damage in less than two minutes in large search spaces without requiring self-diagnosis or pre-specified contingency plans.”⁴⁰ In one scenario, a six-legged robot tasked with adapting to broken legs and motors was asked to evolve a gait in which none of its feet touched the ground—a task the researchers thought impossible to solve. The system, however, subverted the team’s expectations by flipping the robot onto its back and having it walk on its elbows.⁴¹ Such behavior is uncommon or, rather, often physically impossible for most organisms on Earth.

In a similar vein, Watson and colleagues’ work evolving light-following steering behavior in physical robots resulted in locomotion that was both uniquely suited to their hardware setup and unintuitive for human designers.⁴² Derived from Braitenberg’s classic setup,⁴³ the robots employed two wheels, motors, and light sensors; steering behavior was dictated by how much a specific light-sensor reading was translated into driving speed for a specific wheel. Typically, engineers will drive the right wheel proportionally to the left light sensor, and vice versa, to direct such robots toward a goal. While attempting to evolve similar controls with digital evolution, however, Watson and colleagues found that the evolved robots drove toward the light source in surprising ways. “Some *backed up* into the light while facing the dark . . . Others found the source by light-sensitive eccentric spinning” (see Figure 4).⁴⁴ Interestingly enough, not only was the genetic search space related to spinning locomotion much larger than the traditional solution, but it was found that such spinning is actually better suited (with respect to the hardware) to driving at higher speeds, because trajectories can be easily adjusted on the fly.

³⁹ Lehman et al., “Surprising Creativity,” 274.

⁴⁰ Cully et al., “Robots That Can Adapt,” 503.

⁴¹ See the demo video at Evolving AI Lab, “Behavior Performance.”

⁴² Watson et al., “Embodied Evolution.”

⁴³ Braitenberg, *Vehicles*.

⁴⁴ Lehman et al., “Surprising Creativity,” 289.

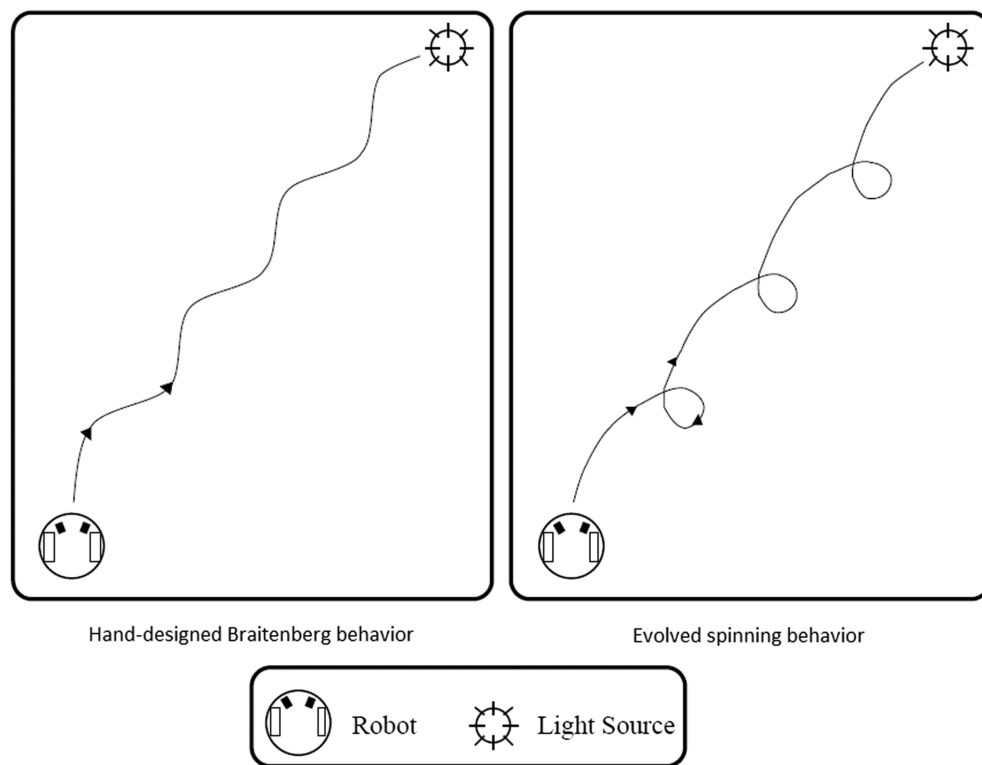


Figure 4 Steering behavior of Watson and colleagues’ light-seeking robot. Left: The locomotion path of a robot employing the traditional Braitenberg approach of proportional left-right steering. Right: The path of a robot using evolved spinning locomotion. Source: Lehman et al., “Surprising Creativity.”

These examples hint at the alternative path dependencies and capabilities evolutionary computational techniques are positioned to produce. Strictly speaking, neither Sims’s creatures nor the unusual evolutionary results presented here were enacted with the explicit intention of creating strange, nonbiomorphic outcomes. However, Sims himself recognized a need for novel ways to architect and construct complex, intelligent systems, noting that “as computers become more powerful, the creation of virtual actors, whether animal, human, or completely unearthly, may be limited mainly by our ability to design them, rather than our ability to satisfy their computational requirements.”⁴⁵ Given the serendipitous and divergent results observed, we suggest that such instances point to a potential space for intentional future exploration. EC has been leveraged as a vehicle for creating more efficient and optimal designs, recreating known designs, and searching through design spaces at paces faster than a human. Through the lens of the bio-xeno spectrum we propose in this work, why not also look to EC as a vehicle for producing fundamentally alien designs?

3.2 Xeno-Design via Machine Learning

If evolutionary approaches allow for forms unconstrained by biological path dependency, then the burgeoning field of GenAI, which combines EC with ML, offers to extend these capabilities even further. One recent example that echoes the spirit of this paper most closely is Tiwary and colleagues’ research on generative visual intelligence (a field the authors dub “GenVI”). In their proposed road map, the authors outline a research agenda that leverages simulation and a combination of genetic and generative techniques to evolve sensing hardware and data-processing methods to craft a new breed of counterfactual visual intelligence. Such an approach, the authors posit, will help humans better understand the “environmental and biological factors that drive the emergence of specific aspects of an animal’s morphology” and create “novel natural imaging systems and behaviors.”⁴⁶ Echoing the sentiments in our framework, the authors reference the potential of GenAI to go beyond biological constraints: “While natural vision is a result of evolution and environmental constraints, we can use GenVI to *generate* new forms of vision.”⁴⁷ Tiwary and colleagues also identify a set of high-potential directions for future investigation: leveraging LLMs to provide a design space (a list of symbols and rules for symbol recombination), searching through the design space via genetic algorithms, reinforcement learning, gradient-based methods, and GenAI-based latent space exploration (e.g., variational autoencoders, generative adversarial networks or GANs, and LLMs), and iterating on the results through simulation, learning, and selection.

⁴⁵ Sims, “Evolving Virtual Creatures.”

⁴⁶ Tiwary et al., “Roadmap for Generative Design.”

⁴⁷ Tiwary et al., “Roadmap for Generative Design,” emphasis in original.

In a follow-up experiment, Kushagra Tiwary and colleagues also demonstrated the use of deep reinforcement learning in single-player games as a method for evolving vision systems in embodied agents. They showed that such simulation could deepen understanding of how environmental factors and specific tasks affect outcomes in eye morphology. In this instance, however, the goal was not to produce novel morphologies, but rather to “recreate the system-level process of vision evolution.”⁴⁸

On the whole, the combining of deep reinforcement learning with advancements in other avenues of ML has proven to be a rather popular approach to developing robotic controls.⁴⁹ However, many such works begin with bioinspired designs (e.g., anthropomorphic bipedal walking system, quadruped, hexapod, etc.) and often aim to optimize factors such as material cost, energy efficiency, or movement speed. As such, these explorations are somewhat tangential to the goals of generative exploration and counterfactual seeking that we emphasize here.

3.3 Xeno-Design via Material-Based Approaches

Finally, outside the realm of simulation, it is worth highlighting that recent work in modular and soft robotics also demonstrates traits that may be helpful in developing the paradigm of xenomorphology. Modular robotics, which concerns mechatronic systems composed of various cooperating units, has already demonstrated numerous creative examples of reconfigurable collectives working together to achieve a common goal—assemblages that we posit are xenomorphic in several aspects. As described by Alattas and colleagues in their review of the space, modular robotics promises to achieve “versatility, robustness, and low cost” by leveraging simple, reproducible modular components that can join together to form diverse assemblies. By applying evolutionary algorithms to this problem space, researchers are achieving configurations that are geared to “allow self-assembly from constituent modules, self-reconfiguration into different functional forms, self-repair to detect errors and recover from failures, and self-reproduce where one system can produce another autonomous functional system.”⁵⁰ The resulting robots—reviewed extensively in works such as Alattas and colleagues and Yim and colleagues⁵¹—exhibit unconventional aesthetics, movement patterns, and affordances that, despite biomimetic origins, inherently diverge from biological norms because of the physical materials involved.

Chin and colleagues’ AuxBots exemplify these characteristics through their use of the expansion and contraction of an auxetic shell to create shape-changing behavior and movement (see Figure 5).⁵² In a similar vein, John Romanishin and colleagues’ M-blocks are composed of self-reconfiguring cubic modules that bond through embedded magnets. Because the individual cubic modules can pivot on any one of twelve edges and contain internal actuators, the resulting observed behavior features various collective-driven obstacle traversal maneuvers, concave transitions, convex transitions, and translations.⁵³ Modules can also come together to form structures, as in Figure 6. While swarm behavior in nature might mirror some of this behavior, the particular manner in which such modular robots complete their tasks is unique in material, aesthetic, and overall locomotion. The wide range of human-designed morphological solutions exhibited by such works point to a vast design space unavailable to natural evolutionary pressures. Continued exploration with the help of, e.g., EC, might open up this space further. For these reasons, we look to work in modular robotics as sources of inspiration when considering the meaning of xenomorphia.

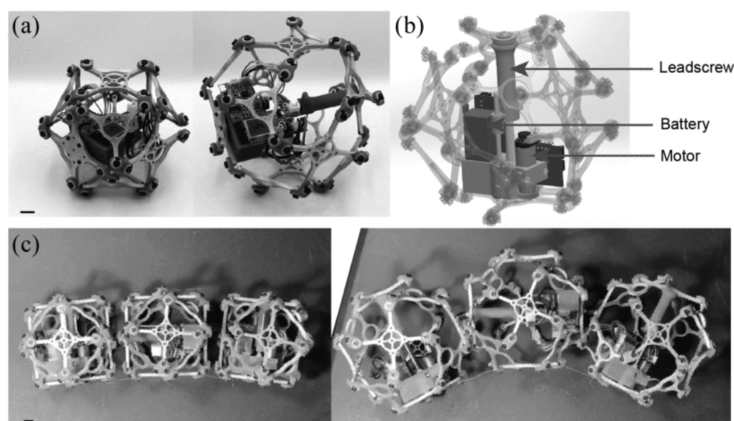


Figure 5 The AuxBots are composed of modules that expand and contract (a) to enable bending and forward motion when connected with wire constraints (c). Their individual make-up is shown in (b). Source: Chin et al., “Flipper-Style Locomotion Through Strong Expanding Modular Robots.”

⁴⁸ Tiwary et al. “What If Eye...,” 3.

⁴⁹ Chen et al., “Deep Reinforcement Learning”; Kalimuthu et al., “Deep Reinforcement Learning”; Luck et al., “Data-Efficient Co-Adaptation.”

⁵⁰ Alattas et al., “Evolutionary Modular Robotics,” 818.

⁵¹ Alattas et al., “Evolutionary Modular Robotics”; Yim et al., “Modular Self-Reconfigurable Robot.”

⁵² Chin et al., “Flipper-Style Locomotion Through Strong Expanding Modular Robots.”

⁵³ Romanishin et al., “M-Blocks.”

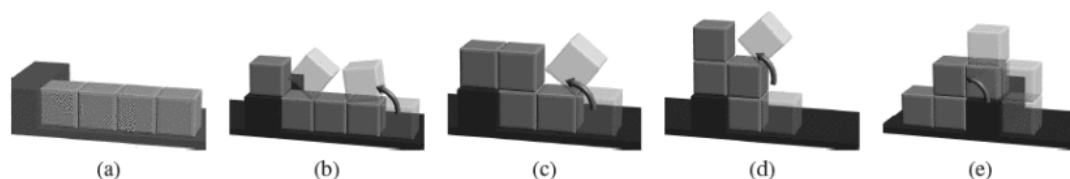


Figure 6 A modeled progression of M-block modules traversing an obstacle (indicated by a black box). Source: Romanishin et al., “M-Blocks.”

Beyond traditional rigid materials, researchers have also begun to investigate the potential of soft, flexible, and biobased materials to enable conformable, shape-changing mechanisms. This results in mechanisms that demonstrate dynamic behaviors and assemblies once considered the exclusive domain of organic tissues. One among many examples is Wani and colleagues’ bionic flytrap, an autonomous liquid-crystal elastomer device that uses optical feedback to trigger a photomechanical, Venus-flytrap-like snapping action (Figure 7).⁵⁴ The device is completely self-contained and does not require electricity or compute or external power, sans natural light. Another soft robot with a rather alien-like gait is Haojian Lu and colleagues’ multilegged millirobot, whose profusion of tapered feet and unassuming appearance seems to occupy a space between that of a caterpillar, starfish, and sentient carpet.⁵⁵ Fabricated out of polydimethylsiloxane (PDMS), hexane, and magnetic particles, the robot’s motion is regulated by the force of an external magnetic field. Figure 8 shows how the movement of a magnetic bar enables the authors to create two distinct gaits. Such work shows how artifacts from soft robotics tend to be recognizable yet alien at the same time.

Yet another bioinspired engineering endeavor, Ren and collaborators’ jellyfish-inspired, soft millirobot⁵⁶ references the fluidic control abilities of *scyphomedusae* ephyra (a type of jellyfish) to craft a soft robot made of magnetic composite elastomer (Figure 9). The final configuration relies on an external oscillating magnetic field to move through space. It can selectively trap and transport objects (see Figure 10), burrow, enhance the mixing of different chemicals in a solution, and generate a concentrated chemical path. Notably, the millibot exhibits the capability to execute five different swimming modes—some more xenomorphic than others: (1) one that attempts to mimic the natural motion of *scyphomedusae* as closely as possible, (2) one characterized by a shorter contraction phase, (3) one with a shorter recovery phase, (4) one with an extra glide phase after contraction, and (5) one with a smaller beating amplitude (see Figure 11). These modes show that nature-inspired design can—through being put in conversation with synthetic and novel materials, and modifications across even a handful of parameters—create something that builds on simple biomimicry to produce something cyborgian. Indeed, in their review of soft robotics for space exploration, Zhang and colleagues identified these jellyfish-inspired devices as potentially useful for exploring “planetary surfaces and even Titan-like planets with lakes.”⁵⁷

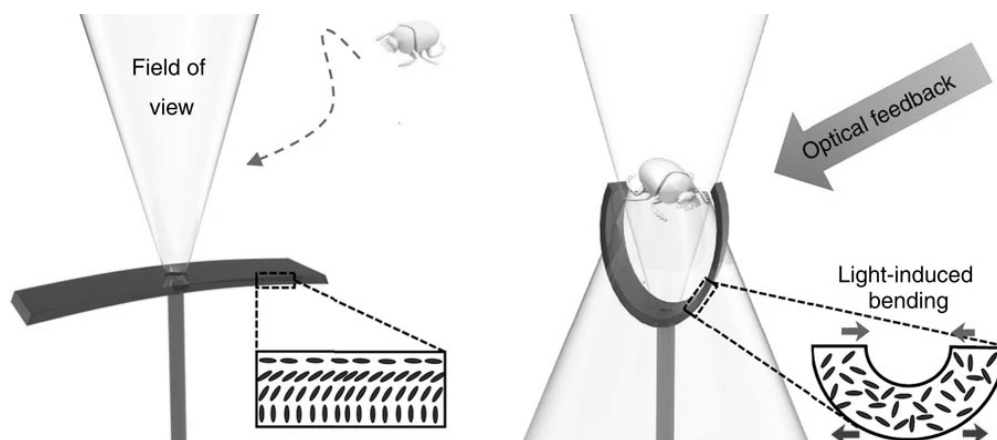


Figure 7 Left: Modeled after the Venus flytrap, the light-triggered artificial flytrap is depicted in its default state. No light is back-reflected to the LCE actuator, shown in red. Right: When an object enters the flytrap’s field of view, it triggers optical feedback to the LCE actuator, which causes the material to bend, the trap to close, and the object to be captured. Source: Wani et al., “Light-Driven Artificial Flytrap.”

⁵⁴ Wani et al., “Light-Driven Artificial Flytrap.”

⁵⁵ Lu et al., “Bioinspired Multilegged Soft Millirobot.”

⁵⁶ Ren et al., “Multi-Functional Soft-Bodied.”

⁵⁷ Zhang et al., “Progress, Challenges, and Prospects,” 11.

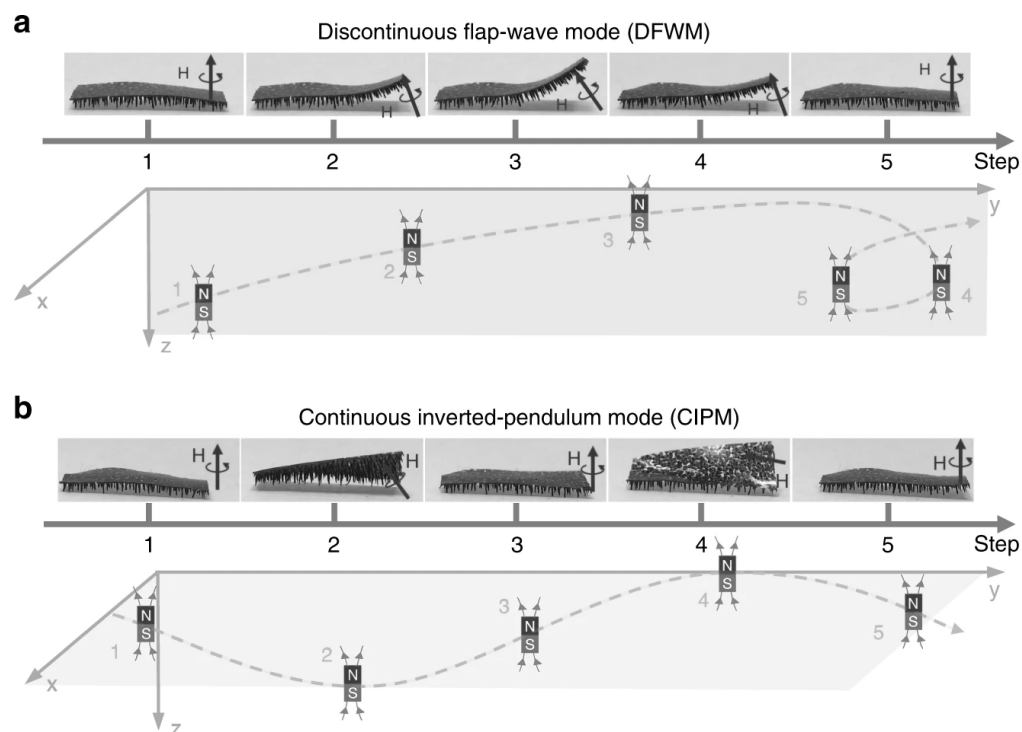


Figure 8 This figure shows the two ways the multilegged millibot can move through space. Below the frames of each locomotive mode is a graph showing along which axes a bar magnet is being manipulated to generate the resulting movement. (a) In the DFWM, the millibot moves according to an “O”-shaped magnetic field on the “y–z” plane. (b) In CIPM, the millibot moves according to an “S”-shaped magnetic field on the “x–y” plane. Source: Lu et al., “Bioinspired Multilegged Soft Millirobot.”

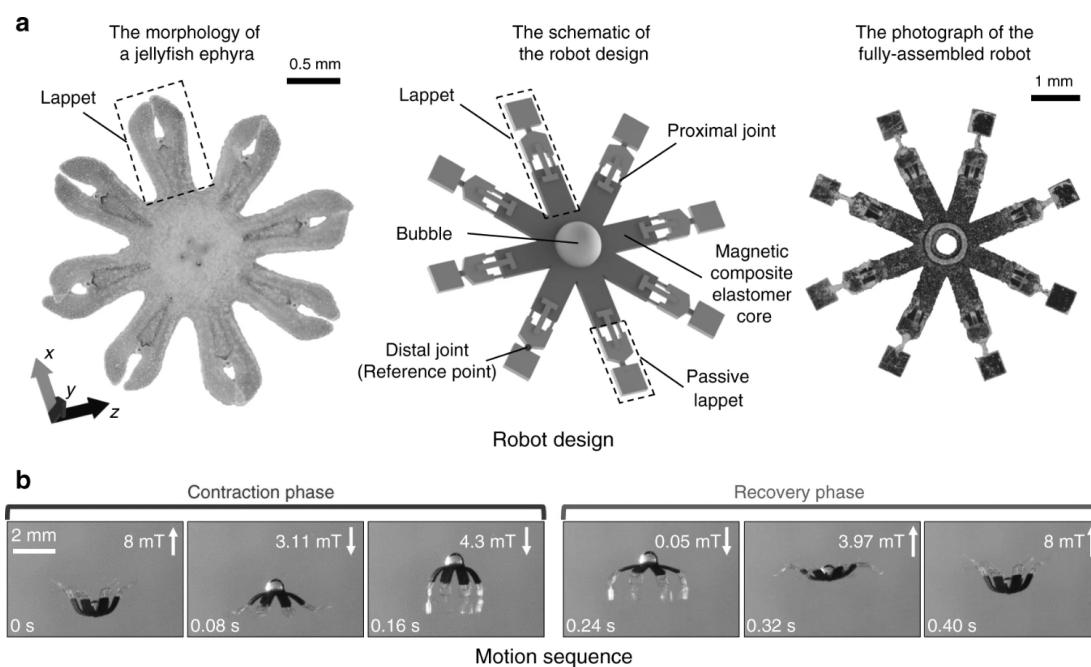


Figure 9 (a) The morphology of scyphomedusae ephyra side by side with the schematic and fabricated millirobot. (b) A motion sequence depicting the millirobot in action, capturing a buoyant bead using the fluid flow around its lappets. Source: Ren et al., “Multi-Functional Soft-Bodied.”

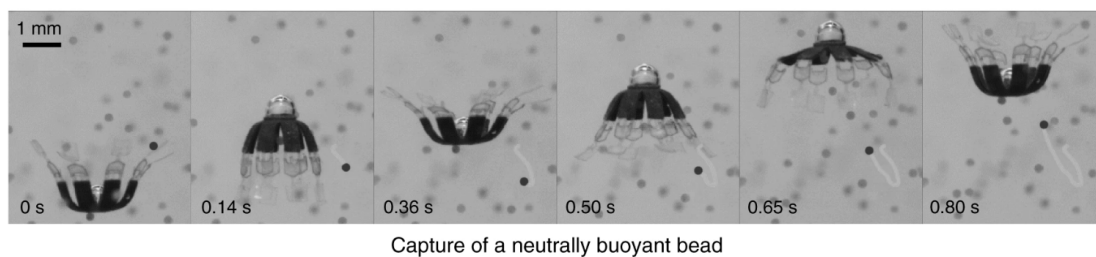


Figure 10 A motion sequence depicting the millirobot in action, capturing a buoyant bead using the fluid flow around its lappets. The red and green annotations indicate the flow of fluid around the millirobot. Source: Ren et al., “Multi-Functional Soft-Bodied.”

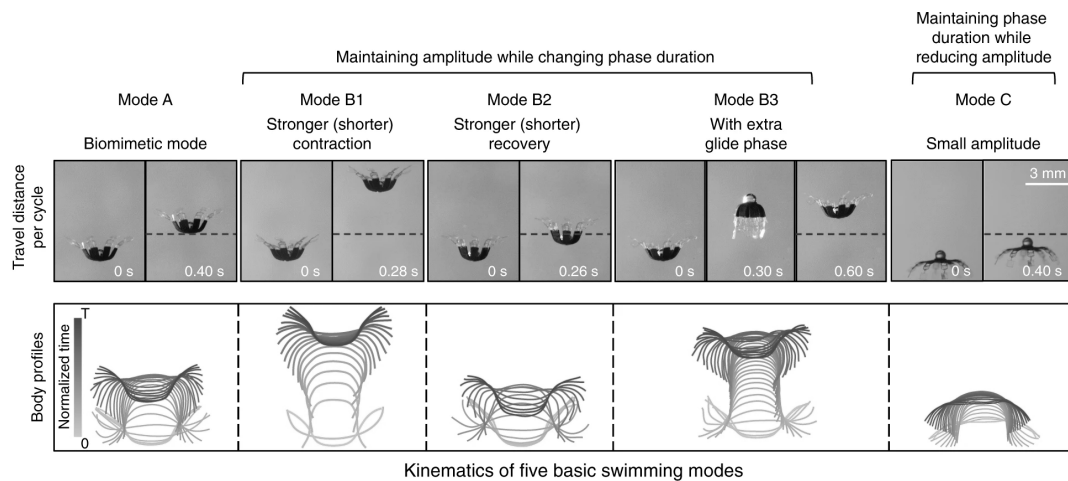


Figure 11 The kinematics of each swimming mode, with the start and end frames of one swimming cycle displayed side by side for each. The dashed red line denotes the final position of the robot when in Mode A to facilitate cross-mode comparison. Source: Ren et al., “Multi-Functional Soft-Bodied.”

3.4 Assembly not Assemblage

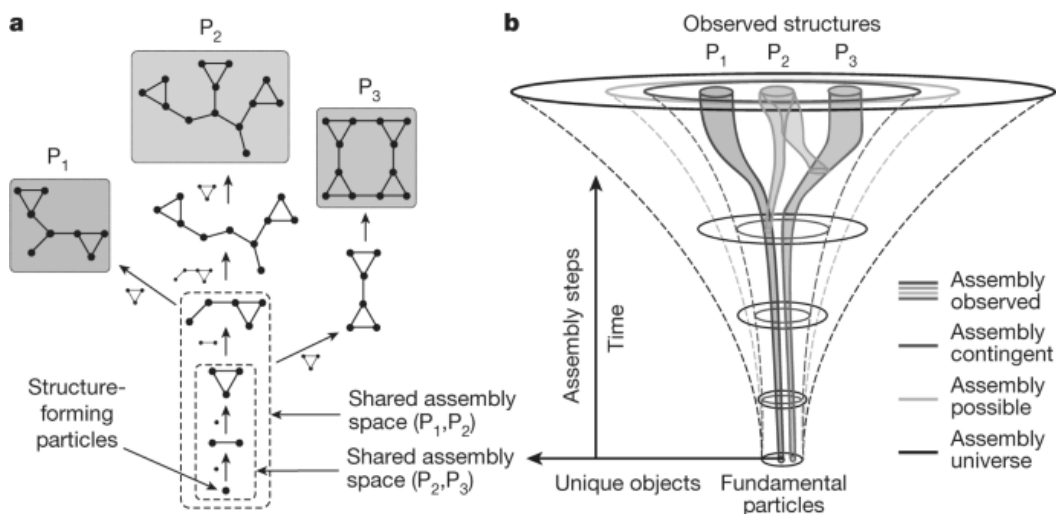


Figure 12 As explained by Sharma et al. in “Assembly Theory Explains: “a, Assembly observed of the three objects shown as graphs (P_1 , P_2 and P_3) with their shared minimal construction process called their ‘joint assembly space’. b, Illustration of the expansion of the assembly universe, assembly possible, assembly contingent and assembly observed (see text for details). Assembly universe has no dynamics and is displayed with assembly steps as the time axis. Note that the figure illustrates their nested structure only, not the relative size of the spaces where each set is typically exponentially larger than the subset.” Source: Sharma et al., “Assembly Theory Explains,” 325.

Where natural evolution produced, among other things, tetrapodal body plans suited to Earth's conditions, techniques incorporating ML, synthetic evolution, and unusual physical materials can explore an almost limitless morphospace. We can imagine assembling entirely new lineages of modular entities—what we call *xenomorphological components*—recombined and reconfigured to produce behaviors and structures devoid of biological precedent. Indeed, it is not uncommon for researchers to tread a middle ground, combining evolutionary algorithms with more traditional algorithmic approaches. As Agoston Eiben and Jim Smith write: “Such hybrid algorithms can often find good (or better) solutions faster than a pure evolutionary algorithm when the additional method searches systematically in the vicinity of good solutions, rather than relying on the more randomized search carried out by mutation.”⁵⁸ Recent advances in machine learning and generative AI are well positioned to accelerate this embracing of that gray area between the nature-inspired and the uniquely machinic.

While such combinatorial experimentation might be loosely described by concepts of “assemblage,” as conceived by contemporary post-Deleuzian readings,⁵⁹ we can gain more precise causal insight by turning to AT. Pioneered by Michael Levin, Lee Cronin, Sara Walker, and collaborators, AT provides a quantitative framework for measuring how many “assembly steps” are needed to form a given object.⁶⁰ Borrowing from molecular assembly theory, AT treats complex biological assemblies as if they were molecular bonds, establishing a minimal path count—an *assembly index*—that captures the structural prerequisites for producing a given entity. Whether an artifact is natural or synthetic, AT tracks the evolutionary (or design) complexity embedded in its form.

Levin's additional work with collaborator McMillen underscores how multiscale architectures define adaptive functionality in biological systems. From cells and tissues up to organs, bodies, and entire swarms, organisms exhibit nested layers of collective intelligence. During embryogenesis, for instance, a blastoderm's cells “agree” on an anatomical fate—say, forming a head versus a tail—through processes of *cellular alignment*. Rather than being a singular vital force, biological unity emerges through compositional engineering across these varied organizational levels.⁶¹ For AT, such coordination can be measured in terms of the assembly steps needed to produce functional outcomes, illuminating how life's remarkable complexity often results from the alignment of subsystems working in concert.

Bringing these threads together suggests a way to design lifelike behaviors or even “new phyla” without relying on genetic inheritance. In doing so, the diagnostic cosmology of AT is extended as a design strategy. By modularizing xenomorphological components and systematically exploring their recombination, researchers could engineer unprecedented forms in silico (or eventually in the physical world), manifesting capacities that surpass biomimetic imitation.

In moving beyond biomimesis, we open a broader morphospace for discovery—one that frames evolution itself as an iterative design process. If the first Cambrian explosion was shaped by environmental conditions and genetic variation, the forthcoming synthetic explosion may be driven by the directed experimentation of AI systems, robotic platforms, and human creativity. Conceiving of evolution as a programmable process—rather than a strictly natural one—promises expansions in both complexity and functionality. In this sense, xenomorphology does more than suggest alien forms: it offers a blueprint for engineering those forms into being, heralding a new era of synthetic morphogenesis on a planetary (and perhaps interplanetary) scale.

4 Xenomorphology as a New Paradigm

The paradigm of generalized biomimesis has thus far exerted considerable influence on experimental thinking in design. While biomimetic approaches aim to replicate the outcomes of terrestrial biology, we propose that a different paradigm is increasingly appropriate. Inspired by AT and Karl Sims's virtual experiments, xenomorphology—and by extension, xenomorphic design—charts a departure from strictly biomorphic principles. Instead of seeking analogies within biological systems, xenomorphology explores forms that evoke true foreignness (*xenos* or ξένος refers to the “strange” or “alien”). By focusing on morphogenetic innovations that do not merely imitate life's evolutionary logic, xenomorphic design envisions shapes and behaviors arising either from the *xenos* of in silico experimentation (as with Sims's work) or, quite literally, from the *xenos* of extraterrestrial environments. To understand this pivot, it is helpful to note that biomorphic design—whether in architecture, robotics, or synthetic biology—relies on the established structures and functions of living organisms. Even when these designs depart from exact replicas (e.g., bipedal robots modeled loosely on human gait), they remain tethered to existing biological archetypes. Xenomorphology, by contrast, proposes forms that, in some way, shape, or form, eschew known biological constraints or evolutionary precedents. Etymologically, “xenomorphic” suggests intrinsically alien morphologies. Popular culture—Ridley Scott's *Alien* foremost among such references—associates “xenomorphs” with disturbing otherness that defies terrestrial norms.

Yet this dichotomy between biomorphism and xenomorphism is not always absolute. In geology, for instance, the terms biomorphic and xenomorphic describe minerals based on their crystallization timeline relative to surrounding structures. Xenomorphic minerals crystallize later,

⁵⁸ Eiben and Smith, 479.

⁵⁹ DeLanda, *Assemblage Theory*; Hayles, “Cognitive Assemblages.”

⁶⁰ Cronin et al., “Assembly Theory.”

⁶¹ McMillen and Levin, *Distributed Intelligence*.

resulting in shapes unconstrained by neighboring formations. This highlights a continuum rather than a strict opposition: biological and nonbiological features often overlap or merge, with each informing the other in unexpected ways.⁶² For design research, acknowledging this fluid interplay can open paths to novel architectures that fuse or move beyond purely organic reference points.⁶³ NASA's TESSERAE (Tessellated Electromagnetic Space Structures for the Exploration of Reconfigurable, Adaptive Environments) project exemplifies how xenomorphic principles can be applied in literal alien environments.⁶⁴ Developed for reconfigurable space architectures, TESSERAE modules are designed to float in microgravity, *quasi-stochastically self-assembling* into desired geometries. Inspired by Roman mosaic tesserae, these tiles can interlock and form larger bases—or be deconfigured and recombined—thus adapting to the vacuum-based constraints of orbit. Crucially, the modules' behavior is not *biologically* derived; they transcend the rigidity of biomorphic forms through their flexible, reconfigurable morphology.



Figure 13 The modular and self-assembly structure of TESSERAE. Source: Artist Rendering of TESSERAE by TU Dortmund – MIT Media Lab.

In virtual contexts, xenomorphological design bypasses biological and physical constraints entirely. Morphologies develop outside of Earthly selection pressures, enabling a catalog of design elements that account for emergent behaviors unrelated to human physiology. Forms might range from enhanced sensory receptors to chemical harvesters or detectors of differential electromagnetism—capabilities that surpass human senses and address digital or off-world requirements. By suspending fitness conditions tied to Earth-based evolution, xenomorphs in silico can evolve traits defying organic intuitions, producing forms and functions that truly challenge our usual design heuristics. Xenomorphological design often involves iterative testing and population-scale experimentation, drawing on combinatorial logics akin to those employed in AlphaFold or reinforcement-learning environments such as Imbue's Avalon.⁶⁵ Instead of a rigid blueprint, an adaptable scaffold supports randomized or algorithmic recombination, allowing xenomorphic responses to emerge on their own terms. This iterative approach could eventually move beyond virtual space—for instance, into a modern “Biosphere 2” setting, where newly designed morphologies adapt in physical but still controlled, semiartificial conditions.

Downstream of xenomorphology lies what we might call *xenobehaviorism*: a lens for analyzing the unique behaviors that alien morphologies engender. Here, AT's concept of an assembly index becomes critical. By tracking the minimal set of “steps” or structural components needed to bring a form into existence, we can interpret how that form might inhabit an alien morphospace and generate novel behaviors unrestricted by Earth's evolutionary lineage. Indeed, xenomorphic forms in silico need not be judged by their potential translation into real-world systems but by the unprecedented behaviors they manifest in virtual or extraterrestrial domains. Fundamentally, true xenomorphology maintains an incommensurability with terrestrial forms. Some influences may, of course, be traced to Earth-based biology, but the foundational structure of xenomorphs should not simply be relabeled biomimicry. This shift compels a reevaluation of how such entities might coexist with humans and with each other—as distinct layers of intelligence and agency that can cooperate or coevolve without collapsing into anthropocentrism. These are alien intelligences, echoing the multiscale “collective intelligence” work of Levin and others, extended to realms beyond the biosphere's known repertoire.

⁶² Kauffman, *Investigations*.

⁶³ Cronin, “Assembly Theory.”

⁶⁴ Ekblaw and Paradiso, “TESSERAE.”

⁶⁵ Albrecht et al., “Avalon.”

In short, where generalized biomimesis anchors design in the familiar terrain of life's historical templates, xenomorphology relinquishes those anchors in pursuit of genuinely novel morphospaces. Whether operating in microgravity, virtual simulation, or any domain unbound by Earthly evolutionary constraints, xenomorphological design invites us to explore—and ultimately, engineer—the truly alien.

5 Conclusion: Speculative Visions for a New Antifragile Xenophylum

For xenomorphological design to fulfill its potential, it must embrace a rigorous scientific and architectural framework. AT offers precisely this: a quantitative approach for charting the structural pathways and “assembly sequences” that lead to novel forms. Coupled with what we might call *xenoarchitectural theory*, researchers can actively anticipate previously unimaginable morphogenetic outcomes. Rather than reproducing life's known forms, this paradigm propels us to explore the radical otherness of xenomorphospace. In doing so, we might cultivate entire environments in which alien intelligences proliferate, challenging established boundaries of morphology, behavior, and human-machine interaction.

The convergence of ideas from Karl Sims, Michael Levin, Lee Cronin, and Sara Walker underscores the unifying role of morphology across the domains of evolution, technology, and intelligence. By quantizing how complex forms come into being—whether through cellular alignment in embryogenesis or modular recombination in virtual simulations—morphology transcends the binary of the organic versus the synthetic. In this sense, xenomorphology emerges not simply as a design novelty but as a structural alternative to biomimesis, realigning the focus of design thinking toward unknown pathways of form.

This perspective resonates with the advent of generative AI, which encourages morphological exploration in its architectures. The most speculative instances of xenomorphological experimentation may initially unfold in silico, but their impact extends to physical infrastructures and cultural imaginaries. Over time, these virtual experiments filter into how we perceive and construct our environments, subtly redefining our relationship to both the biosphere and a nascent xenosphere.

Crucially, it may take adversarial events or stressors to catalyze the emergence of progressively antifragile xenomorphic forms. Because these forms are engineered with artificial responsiveness and flexibility, they adapt more quickly than their biological counterparts. By deploying large-scale simulations and iterative testing—akin to Sims's evolutionary software or Imbue's Avalon environment—designers can accelerate the discovery of xenomorphic responses. When promising new behaviors surface, they can be refined and scaled further, eventually migrating into physical test beds reminiscent of Biosphere 2.

This antifragile imperative—where morphology itself becomes the wellspring for evolving behaviors—reflects a design philosophy attuned to a complex, rapidly shifting world. In reframing evolution as a synthetic and modular process, we glimpse the birth of a new xenophylum of forms adapted to alien intelligences and nonbiological morphospaces. Pushing beyond the constraints of terrestrial body plans, xenomorphology embraces the alien in all its unsettling potential. By prioritizing antifragility and dynamic adaptation, we can seed resilient, exploratory, and provocative designs that transcend the boundaries of the biosphere—marking a transformative milestone in both the theory and practice of design.

Sometimes, an adversarial event can induce an unexpected, xenomorphic response. In this case, acquired xenomorphic behaviors might lead to greater resilience and increased proliferation in various contexts. This insight underscores the need to rethink design frameworks through the lens of a new xenomorphic paradigm, which emphasizes adaptability and innovative responses to challenges. By adopting this paradigm, designers can cultivate more resilient systems that are better equipped to navigate complexity and engineer this new Cambrian explosion.

In this framework, the antifragile response mediates between the morphology (which begets the morphospace of new behaviors) and the emergence of the xenobehavior itself. Without some adversarial stimulus, in other words, a potential xenobehavior may never emerge. Thus, as with Levin's argument for problem-solving capabilities being distributed beyond the bounds of the individual agent, the possibilities of xenomorphology either can be left to chance, to emerge within adversarial encounters, or else such antifragile stimuli can be experimented with. The stimuli that might solicit such a response may be adversarial but are inherently unpredictable and difficult to predetermine. Thus, as with Sims's virtual evolved creatures, experimenting with such emergence benefits from a scaling of experimental frequency.

A small number of xenomorphological components might be recombined either randomly or according to a specific algorithmic logic based on the “multi-sequence alignments” of projects such as AlphaFold.⁶⁶ To establish their postcombinatorial morphospace for potentially novel behaviors, these assembled xenomorphs must be tested in potentially adversarial environments. Promising novel behaviors might be iterated upon.

⁶⁶ Jumper *et al.*, “Highly Accurate Protein Structure,” 583.

Bibliography

- Alattas, Reem J., Sarosh Patel, and Tarek M. Sobh. "Evolutionary Modular Robotics: Survey and Analysis." *Journal of Intelligent & Robotic Systems* 95 (2019): 815–28. <https://doi.org/10.1007/s10846-018-0902-9>.
- Albrecht, Joshua, Abraham J. Fetterman, Bryden Fogelman, et al. "Avalon: A Benchmark for RL Generalization Using Procedurally Generated Worlds." Preprint, arXiv, October 24, 2022. <https://doi.org/10.48550/arXiv.2210.13417>.
- Arias-Montano, Alfredo, Carlos A. Coello Coello, and Efrén Mezura-Montes. "Multiobjective Evolutionary Algorithms in Aeronautical and Aerospace Engineering." *IEEE Transactions on Evolutionary Computation* 16, no. 5 (2012): 662–94. <https://doi.org/10.1109/TEVC.2011.2169968>.
- Bacon, Francis. *Novum Organum*, edited by Joseph Devey. P. F. Collier & Son, 1902. First published in 1620.
- Baker, Ellie, and Margo Seltzer. 1993. *Evolving Line Drawings*. Harvard Computer Science Group Technical Report TR-21-93. Harvard University. <https://dash.harvard.edu/server/api/core/bitstreams/7312037d-ccfe-6bd4-e053-0100007fdf3b/content>.
- Barros, Rodrigo Coelho, Márcio Porto Basgalupp, Andre C. P. L. F. De Carvalho, and Alex A. Freitas. "A Survey of Evolutionary Algorithms for Decision-Tree Induction." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, no. 3 (2011): 291–312. <https://doi.org/10.1109/TSMCC.2011.2157494>.
- Bates, David. *Artificial Intelligence and the Human: History and Future*. MIT Press, 2024.
- Belfiore, Elizabeth. "A Theory of Imitation in Plato's Republic." *Transactions of the American Philological Association* (1974–) 114 (1984): 121–46. <http://www.jstor.org/stable/284143>.
- Benyus, Janine. *Biomimicry: Innovation Inspired by Nature*. Harper Perennial, 1997.
- Biles, John. "GenJam: A Genetic Algorithm for Generating Jazz Solos." In *ICMC Proceedings 1994*. International Computer Music Association, 1994.
- Blanchard, Pierre. *The Dynamics of Mimesis: Cultural Representations and Symbolic Forms*. Cambridge University Press, 1997.
- Braidotti, Rosi. *Posthuman Knowledge*. Polity, 2019.
- Braitenberg, Valentino. *Vehicles: Experiments in Synthetic Psychology*. MIT Press, 1986.
- Brynjolfsson, Erik, and Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W.W. Norton & Company, 2014.
- Castoriadis, Cornelius. *The Imaginary Institution of Society*. Cambridge: MIT Press, 1975.
- Cerella, Antonio. "Until the End of the World: Girard, Schmitt and the Origins of Violence." *Journal of International Political Theory* 11, no. 1 (2015): 42–60. <https://doi.org/10.1177/1755088214555457>.
- Chen, Ci, Pingyu Xiang, Jingyu Zhang, Rong Xiong, Yue Wang, and Haojian Lu. "Deep Reinforcement Learning Based Co-Optimization of Morphology and Gait for Small-Scale Legged Robot." *IEEE/ASME Transactions on Mechatronics* 29, no. 4 (2023): 2697–2708. <https://doi.org/10.1109/TMECH.2023.3330427>.
- Cheney, Nick, Robert MacCurdy, Jeff Clune, and Hod Lipson. "Unshackling Evolution: Evolving Soft Robots with Multiple Materials and a Powerful Generative Encoding." *SIGEVOlution* 7, no.12 (August 2014): 11–23. <https://doi.org/10.1145/2661735.2661737>.
- Chin, Lillian, Max Burns, Gregory Xie, and Daniela Rus. "Flipper-Style Locomotion Through Strong Expanding Modular Robots." *IEEE Robotics and Automation Letters* 8, no. 2 (2023): 528–35. <https://doi.org/10.1109/LRA.2022.3227872>.

- Corucci, Francesco, Marcello Calisti, Helmut Hauser, and Cecilia Laschi. "Novelty-Based Evolutionary Design of Morphing Underwater Robots." In *GECCO '15: Proceedings of the 2015 annual conference on Genetic and Evolutionary Computation*, edited by Sara Silva. Association for Computing Machinery, 2015.
- Cully, Antoine, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. "Robots That Can Adapt Like Animals." *Nature* 521, no. 7553 (2015): 503–7. <https://doi.org/10.1038/nature14422>.
- Darwin, Charles. *On the Origin of Species by Means of Natural Selection*. D. Appleton and Company, 1861.
- DeLanda, Manuel. *Assemblage Theory*. Edinburg University Press, 2016.
- Dupuy, Jean-Pierre. *The Mechanization of the Mind: On the Origins of Cognitive Science*. Princeton University Press, 1992.
- Dumouchel, Paul. *The Barren Sacrifice: An Essay on Political Violence*. Michigan State University Press, 2014.
- Eiben, Agoston E., and Jim Smith. "From Evolutionary Computation to the Evolution of Things." *Nature* 521, no. 7553 (2015): 476–82. <https://doi.org/10.1038/nature14544>.
- Ekblaw, Ariel, and Joseph Paradiso. "TESSERAE: Self-Assembling Shell Structures for Space Exploration." In *Proceedings of the Annual Symposium of the International Association of Shell and Spatial Structures: Extra Planetary Architecture*. IASS, 2018.
- Evolving AI Lab. "A Behavior Performance Map Containing Many Different Types of Walking Gaits." May 1, 2017. YouTube, 2:12. <https://www.youtube.com/watch?v=H6OB1E8NsLw&list=PL5278ezwmoxQODgYB0hWnC0-Ob09GZGe2&t=108s>.
- Floridi, Luciano. *The Ethics of Information*. Oxford University Press, 2011.
- Frank, Adam. "The Coming Second Copernican Revolution." *Noema Magazine*, October 15, 2024. <https://www.noemamag.com/the-coming-second-copernican-revolution/>.
- Geertz, Clifford. *Works and Lives: The Anthropologist as Author*. Stanford University Press, 1988.
- Girard, René. *Violence and the Sacred*. Johns Hopkins University Press, 1977.
- Hayles, N. Katherine. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. University of Chicago Press, 1999.
- . "Cognitive Assemblages: Technical Agency and Human Interactions." *Critical Inquiry* 43, no. 1 (2016): 32–55. <https://doi.org/10.1086/688293>.
- Hui, Yuk. *Recursivity and Contingency*. Rowman & Littlefield, 2023.
- Jumper, John, Richard Evans, Alexander Pritzel, et al. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (2021): 583–89. <https://doi.org/10.1038/s41586-021-03819-2>.
- Kalimuthu, Manivannan, Abdullah Aamir Hayat, Thejus Pathmakumar, Mohan Rajesh Elara, and Kristin Lee Wood. "A Deep Reinforcement Learning Approach to Optimal Morphologies Generation in Reconfigurable Tiling Robots." *Mathematics* 11, no. 18 (2023): 3893. <https://doi.org/10.3390/math11183893>.
- Kapp, Ernst. *Elements of a Philosophy of Technology: On the Evolutionary History of Culture*. University of Minnesota Press, 2018.
- Kauffman, Stuart A. *Investigations*. Oxford University Press, 2000.
- Khemka, Namrata, Scott Novakowski, Gerald Hushlak, and Christian Jacob. "Evolutionary Design of Dynamic Swarmscapes." In *GECCO '08: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, edited by Maarten Keijzer. Association for Computing Machinery, 2008.

- Lehman, Joel, Jeff Clune, Dusan Misevic, et al. "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities." *Artificial Life* 26, no. 2 (2020): 274–306. https://doi.org/10.1162/artl_a_00319.
- Lei, Zhenyu, Shangce Gao, Zhiming Zhang, MengChu Zhou, and Jiujun Cheng. "MO4: A Many-Objective Evolutionary Algorithm for Protein Structure Prediction." *IEEE Transactions on Evolutionary Computation* 26, no. 3 (2021): 417–30. <https://doi.org/10.1109/tevc.2021.3095481>.
- Lessin, Dan, and Sebastian Risi. "Soft-Body Muscles for Evolved Virtual Creatures: the Next Step on a Bio-Mimetic Path to Meaningful Morphological Complexity." In *Proceedings of the European Conference on Artificial Life (ECAL) 2015*, edited by Paul Andrews, Leo Caves, René Doursat, et al. MIT Press, 2015.
- Lu, Haojian, Mei Zhang, Yuanyuan Yang, et al. "A Bioinspired Multilegged Soft Millirobot that Functions in Both Dry and Wet Conditions." *Nature Communications* 9, no. 3944 (2018): 3944. <https://doi.org/10.1038/s41467-018-06491-9>.
- Luck, Kevin Sebastian, Heni Ben Amor, and Roberto Calandra. "Data-Efficient Co-Adaptation of Morphology and Behaviour with Deep Reinforcement Learning." In *Proceedings of the Conference on Robot Learning*, edited by Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura. PMLR 100, November 2019. <https://proceedings.mlr.press/v100/luck20a.html>.
- McLuhan, Marshall. *Understanding media: The extensions of man*. McGraw-Hill, 1964.
- McMillen, Colin, and Michael Levin. *Distributed Intelligence and Morphogenesis: The Multiscale Architecture of Biological Problem Solving*. MIT Press, 2024.
- Mill, John S. "On Nature." In *Nature, The Utility of Religion and Theism*. Longmans, Green, Reader, and Dyer, 1875.
- Mormino, Gianfranco. *Per una teoria dell'imitazione*. Raffaello Cortina Editore, 2016.
- Moore, George Edward. *Principia Ethica*. 2nd ed. Cambridge University Press, 2008. First published in 1903.
- Nolfi, Stefano, and Dario Floreano. *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. MIT Press, 2000. <https://doi.org/10.7551/mitpress/2889.001.0001>.
- Palaver, Wolfgang. *René Girard's Mimetic Theory*. Michigan State University Press, 2013.
- Pawlyn, Michael. *Biomimicry in Architecture*. RIBA Publishing, 2019.
- Polites, Nikolas. *Sustainable Design: Biomimetic Approaches to Innovation*. Columbia University Press, 2019.
- Ren, Ziyu, Wenqi Hu, Xiaoguang Dong, and Metin Sitti. "Multi-Functional Soft-Bodied Jellyfish-Like Swimming." *Nature Communications* 10, no. 2703 (2019): 2703. <https://doi.org/10.1038/s41467-019-10549-7>.
- Romanishin, John W., Kyle Gilpin, and Daniela Rus. "M-Blocks: Momentum-Driven, Magnetic Modular Robots." In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013. <https://doi.org/10.1109/IROS.2013.6696971>.
- Sawada, Junji, Satoru Hiwa, and Tomoyuki Hiroyasu. "Evolutionary Generative Design: Integrating Machine Learning and Evolutionary Computation for Automated Design Space Exploration." Preprint, TechRxiv, January 31, 2025. <https://doi.org/10.36227/techrxiv.173834874.46747234/v1>.
- Sharma, Abhishek, Dániel Czégel, Michael Lachmann, Christopher P. Kempes, Sara I. Walker, and Leroy Cronin. "Assembly Theory Explains and Quantifies Selection and Evolution." *Nature* 622 (2023): 321–28. <https://doi.org/10.1038/s41586-023-06600-9>.
- Sims, Karl. 1994. "Evolving Virtual Creatures." In *SIGGRAPH '94: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*. Association for Computing Machinery, 1994. <https://doi.org/10.1145/192161.192167>.

- Stiegler, Bernard. *Technics and Time, 1: The Fault of Epimetheus*. Stanford University Press, 2018.
- Telikani, Akbar, Amirhessam Tahmassebi, Wolfgang Banzhaf, and Amir H. Gandomi. “Evolutionary Machine Learning: A Survey.” *ACM Computing Surveys (CSUR)* 54, no. 8 (2021): 161. <https://doi.org/10.1145/3467477>.
- Tiwary, Kushagra, Tzofi Klinghoffer, Aaron Young, et al. “A Roadmap for Generative Design of Visual Intelligence.” An MIT Exploration of Generative AI, September 2024. <https://doi.org/10.21428/e4baedd9.d2a03144>.
- Tiwary, Kushagra, Aaron Young, Zaid Tasneem, et al. “What if Eye...? Computationally Recreating Vision Evolution.” Preprint, arXiv, January 25, 2025. <https://doi.org/10.48550/arXiv.2501.15001>.
- Vincent, Julian F. V., Olga A. Bogatyreva, Nikolaj R. Bogatyrev, Adrian Bowyer, and Anja K. Pahl. “Biomimetics: Its Practice and Theory.” *Journal of the Royal Society Interface* 3, no. 9 (2006): 471–82. <https://doi.org/10.1098/rsif.2006.0127>.
- Wani, Owies M., Hao Zeng, and Arri Priimagi. “A Light-Driven Artificial Flytrap.” *Nature Communications* 8 (2017): 15546. <https://doi.org/10.1038/ncomms15546>.
- Watson, Richard A., Sevan G. Ficici, and Jordan B. Pollack. “Embodied Evolution: Distributing an Evolutionary Algorithm in a Population of Robots.” *Robotics and Autonomous Systems* 39, no. 1 (2002): 1–18. [https://doi.org/10.1016/S0921-8890\(02\)00170-7](https://doi.org/10.1016/S0921-8890(02)00170-7).
- Widera, Paweł, Jonathan M. Garibaldi, and Natalio Krasnogor. “GP Challenge: Evolving Energy Function for Protein Structure Prediction.” *Genetic Programming and Evolvable Machines* 11 (2010): 61–88. <https://doi.org/10.1007/s10710-009-9087-0>.
- Yao, Xin. “Evolving Artificial Neural Networks.” *Proceedings of the IEEE* 87, no. 9 (1999): 1423–47.
- Yim, Mark, Wei-Min Shen, Behnam Salemi, Daniele Rus, et al. “Modular Self-Reconfigurable Robot Systems.” *IEEE Robotics & Automation Magazine* 14, no. 1 (2007): 43–52.
- Zhang, Yongchang, Pengchun Li, Jiale Quan, Longqiu Li, Guangyu Zhang, and Dekai Zhou. “Progress, Challenges, and Prospects of Soft Robotics for Space Applications.” *Advanced Intelligent Systems* 5 (2023): 2200071. <https://doi.org/10.1002/aisy.202200071>.



3 Organs Without Bodies

Where does mere information processing end and active cognition begin? As artificial intelligence advances, the boundary between these two states becomes increasingly ambiguous. Evolutionary biology offers a valuable perspective: Historically, sensory organs such as eyes have played a critical role in driving the development of brains, emphasizing that cognition emerges from sensory capacities. Thinking, therefore, is inherently tied to sensing—an insight equally pertinent to artificial sensing and intelligence.

The emergence and proliferation of new, cognitively active forms of intelligence necessitates a fundamental reimagining and reengineering of the relationship between intelligence and embodiment. Material substrates that inherently possess cognitive properties, such as neural tissue, are being integrated into innovative technological assemblages. Simultaneously, forms of embodiment traditionally associated primarily with sensory roles are evolving to actively participate in cognitive processes, transcending their original function of merely sensing environmental information.

Such multiplicities in cognition and embodiment should not be viewed as anomalies; rather, they reflect the intrinsic plurality already present within biological systems. The brain itself exemplifies this multiplicity, with cortical columns concurrently negotiating diverse aspects of experience in both integrative and divergent ways. Extending this principle beyond individual organisms, cognition similarly manifests as a dynamic interplay among multiple embodied entities, each contributing uniquely to the broader cognitive landscape.

These projects examine these transformative developments, exploring the philosophical and practical implications of redefining cognition in relation to novel modes of embodiment. By appreciating the pluralistic nature of cognitive processes, they aim to expand our understanding of how emerging forms of artificial intelligence and sensory integration reshape the fundamental boundaries of cognition itself.

3a *Organoid Array Computing*

When exploring potential substrates for computation, the human brain naturally stands out as a highly sophisticated example. Biological neural networks and the evolutionary phenomenon of cephalization have long been intertwined, yet it remains unclear whether their coupling is indispensable for cognition or merely one evolutionary pathway among many. Recent advances in brain organoid research suggest intriguing alternatives. Brain organoids—laboratory-grown clusters of neural tissue—demonstrate remarkable cognitive capacities, including responding to stimuli, generating measurable brain waves, controlling rudimentary robotic systems, and even performing tasks such as playing *Pong*.

This research prompts critical questions about the materiality of intelligence itself. Brains, as substrates, clearly possess inherent plasticity suited to artificialized forms of computation, suggesting untapped potential within the broader landscape of “learning matter.” Rather than programming artificial intelligence, what new possibilities emerge when we instead grow it organically?

Yet, the human brain’s complexity arises largely from its extensive neural networks and division of labor across specialized regions. Could similar complexity emerge from interconnected networks of brain organoids? This project explores this idea, imagining the cultivation of organoid networks capable of mutual communication, hypothesizing that such interconnected systems will yield increasingly sophisticated cognitive behaviors. These interactions may foster evolutionary-like dynamics among organoids, with certain units potentially developing greater adaptability and efficiency than others.

Should this scenario materialize, we would indeed have grown a unique form of artificial intelligence—distinct yet comparable to silicon-based systems. However, important questions remain unresolved: What specific advantages might this organically derived intelligence hold over traditional computational substrates? This project probes these boundaries, illuminating novel paths forward in the ongoing quest to understand and engineer intelligence.

3b *Cognition With and Beyond the Brain*

Cognition is not confined solely to the brain; it emerges dynamically through interactions extending across and beyond the physical boundaries traditionally associated with thought. Given this broader conceptualization, it stands to reason that technological augmentation of cognition should similarly extend beyond cerebral confines. Historically, technological enhancements of the human body have predominantly targeted sensory mediation, limiting augmentation to the refinement of external inputs. However, recognizing that cognition permeates the entirety of embodied experience opens new possibilities for integrative augmentation.

Indeed, numerous species exhibit decentralized cognitive capabilities distributed throughout their bodies, raising intriguing questions about the potential for creating analogous technological transduction layers within humans. What forms could such a layer take, and how might it expand our cognitive horizons?

This paper provides a philosophical foundation for emerging and existing forms of artificial sensory and somatosensory augmentation. Bridging philosophical inquiry with contemporary technological developments, the project draws on Bernard Stiegler’s philosophical explorations of “endosomatization”—or, as adapted here, “intrasomatization”—to analyze advanced epidermal media and computational frameworks embedded directly in bodily tissues.

By integrating these philosophical perspectives with cutting-edge engineering, this paper explores how cognition might be technologically extended across bodily surfaces, fundamentally transforming the interface between human bodies, perception, and thought processes. Ultimately, this interdisciplinary approach lays the conceptual groundwork for understanding and developing novel augmentations that acknowledge and leverage the distributed nature of cognition.



Organoid Array Computing

The Design Space of Organoid Intelligence

Jenn Leung
Uni. of the Arts
London

Chloe Loewith
University of
Cambridge

Ivar Frisch
Utrecht
University

Abstract

In this paper, we explore the artificialization and networking of biological matter via brain organoids—three-dimensional, stem-cell-derived structures that recapitulate aspects of human brain architecture and function. These organoids serve as a platform for investigating the emergent properties of biological neural networks and the potential for developing an in-vitro to in-silico cognitive architecture. Our research addresses the burgeoning field of organoid intelligence (OI), wherein biological substrates are interfaced with computational systems, providing an adaptive framework for embodied computation. A common distinction between software and hardware in the field of biocomputing assumes DNA as software and cells as hardware. By evolving through biochemical and physical signaling feedback, organoids challenge this dichotomy. OI integrates both, enabling biological systems to move along the continuum from software to hardware into a multiscale machine. We begin by examining the current interfacing technologies that enable the connection between organoids and digital systems, evaluating the proof-of-concept studies that have laid the groundwork for OI applications. This analysis includes a critical assessment of the existing practical and technical limitations that hinder the realization of scalable OI. We then propose design strategies aimed at overcoming these obstacles, emphasizing the need for a nested approach to experimental design. New permutations enable the iterative development of OI modules, facilitating the integration and application of polycomputational neural assemblies. The design space of OI focuses on the growing dimensions and analysis of inputs, outputs, interfaces, and frameworks across multiple scales. We posit that the design of OI is less an act of top-down design and more a process of guided evolution, wherein higher-order cognitive functions emerge organically from the intricate interplay of lower-level biochemical substrates. Through this, we speculate on how higher-order functions can emerge from networking biological matter from embedded substrates “downstream”. Our research aims to uncover new dimensions in the information-processing capabilities of OI, positioning OI as a novel form of AI.

Keywords

artificial intelligence; human brain organoids (HBOs); organoid intelligence (IO); interface; biological neural network; hybridized entities; biocomputing

1 Introduction

The discovery of new materials and functional substrates continually reshapes novel approaches in computing and provides “new means of acting on and interpreting the world.”¹ Recent advancements in stem cell research have opened groundbreaking avenues for utilizing human neuronal tissue as a substrate for unconventional neural networks. At the forefront of this research are brain organoids: complex structures derived from stem cells that develop functional neural networks, enabling the artificialization of biological matter and the networking of growing and thinking matter.²

Organoids, when interfaced with artificial intelligence through multielectrode arrays (MEAs), present a compelling framework for instantiating novel modes of computing and neural network architectures. This paper explores the design space of organoid intelligence (OI), traversing the continuum from organoids as “software” emerging from lower-level biological substrates to their role as “hardware” supporting higher-order computational processes.

If the word *computer* refers to any physical object that can implement any computable function, then biological brains—and, by extension, human brain organoids (HBOs)—are literally computers.³ Viewing biological neural tissues as computational entities enables us to move beyond the conception of computers as externalized brains, instead considering artificialized brain models as computers themselves. Further, as current developments in artificial intelligence point toward a “hardware bottleneck issue,” this position enables us to depart from neuromorphic designs in silicon and to examine the inherent computational power of living neuronal tissues as a potential solution.⁴

We propose viewing the design space of OI as a “polycomputational” neural assembly, integrating hardware, software, organoids, and chemical substrates into a cohesive computational framework that couples in-vitro biological systems with in-silico computational frameworks. In the following sections, the paper will elucidate and expand on the design space of OI, delineating the various dimensions crucial to experiment design in this emerging field. We will explore the building blocks of OI experiments, considering inputs, outputs, interfaces, and frameworks across multiple scales. By mapping this design space, we seek to provide a comprehensive foundation for future OI experimentation and its potential applications in artificial intelligence.

2 Background on Organoids

Organoids are 3D tissue cultures derived from stem cells that model the structures and functions of an organ.⁵ The discovery of organoids began with Hans Clevers’s groundbreaking work, in which organoids were first grown in lab petri dishes using stem cells from patients’ small intestines.⁶

Advancements in stem cell research and tissue cultures have blurred the lines between biological and artificial systems. Key developments include the creation of chimeric embryos, which combine genetic material from different organisms, and synthetic tissue cultures that challenge the primacy of DNA in determining biological outcomes.⁷ These breakthroughs not only expand our understanding of biological plasticity but also pave the way for innovative research in regenerative medicine, organ transplantation, and the study of human evolution and neurodevelopment.⁸

Today, stem cell cultures can be programmed and exposed to specific environmental factors to functionally model various organ sections.⁹ HBOs, also referred to as cerebral organoids, are grown from human pluripotent stem cells that can mimic aspects of the architecture and functionality of the human brain (see Figure 1).¹⁰ HBOs are traditionally utilized for studying human brain development, modeling neurological diseases, testing drug efficacy, and other assessments of neurodevelopmental processes that would otherwise present ethical and practical constraints associated with human brain research.¹¹

Although HBOs have the ability to replicate certain aspects of brain structure and function, they are currently limited in their scale and complexity. Without solving significant gaps in vascularization and interorganoid communication, organoids remain a minimum “working model of some of the circuitry resident in a living, functioning human brain.”¹²

¹ Beaulieu et al., “Refractive Computation,”

² Smirnova et al., “Organoid Intelligence: New Frontier”

³ Richards and Lillicrap, “Brain-Computer Metaphor.”

⁴ Mencattini, “Assembloid Learning.”

⁵ Smirnova et al., “Organoid Intelligence: New Frontier”

⁶ Sato et al., “Single Lgr5 Stem Cells.”

⁷ Blakemore, “Human-Pig Hybrid”; Kruszelnicki, “Mouse with Human Ear”; Tissue Culture & Art Project, “Crude Matter.”

⁸ (Sun et al., “Applications of Brain Organoids” 2021; Chen et al., “Human Brain Organoids.” 2019

⁹ Fernandes, “Organoids as Complex (Bio)Systems.”

¹⁰ Takahashi and Yamanaka, “Induction of Pluripotent Stem Cells”; Baldassari, “Brain Organoids.”

¹¹ Sun et al., “Applications of Brain Organoids.”

¹² Goldman, “Assembloid Models”; Smirnova et al., “Organoid Intelligence: New Frontier.”

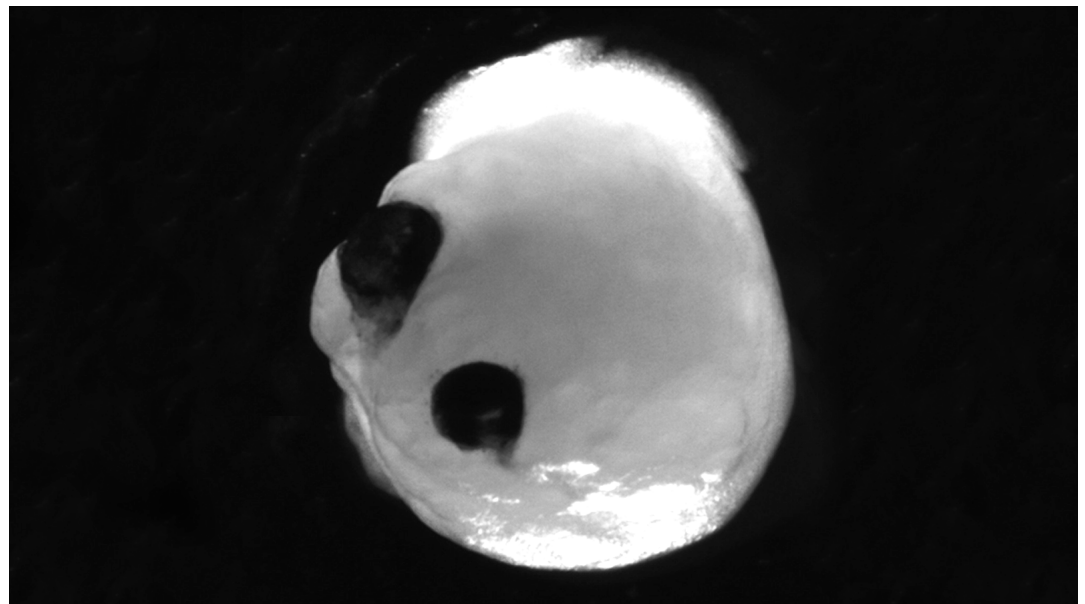


Figure 1 Sixty-day-old organoids with bilaterally symmetric pigmented optic vesicles.¹³

2.1 The Concept of Organoid Intelligence: In-vitro to In-silico Interface

An organoid in a petri dish exists in an in-vitro state, isolated from external inputs and without perceivable outputs. However, HBOs have now been interfaced with artificial intelligence, creating an interconnected system where AI serves as an analytical tool to process high-dimensional data from these biological structures.¹⁴ This is referred to as “organoid intelligence” (OI), a term first introduced in the article “Organoid Intelligence (OI): The New Frontier in Biocomputing and Intelligence-in-a-Dish.”¹⁵

OI is the hybridization of biological computing with machine interface technologies. This integration enables us to virtually embody organoids, transitioning them from an in-vitro to an in-silico “in computer” instance. The physical organoid interface includes three central components: an HBO, an MEA, and a microfluidic platform.¹⁶ Each element plays a crucial role in creating a functional and interactive bio-electronic interface (Figure 2).

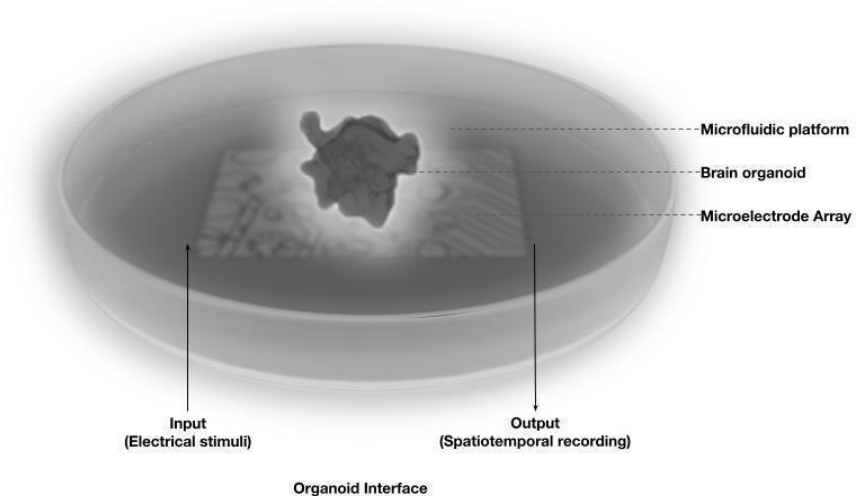


Figure 2 Typical interface for OI. Figure designed by Jenn Leung.

¹³ Gabriel et al., “Human Brain Organoids.” 2021

¹⁴ Smirnova et al., “Organoid Intelligence: New Frontier.”

¹⁵ Smirnova et al., “Organoid Intelligence: New Frontier.” 2023

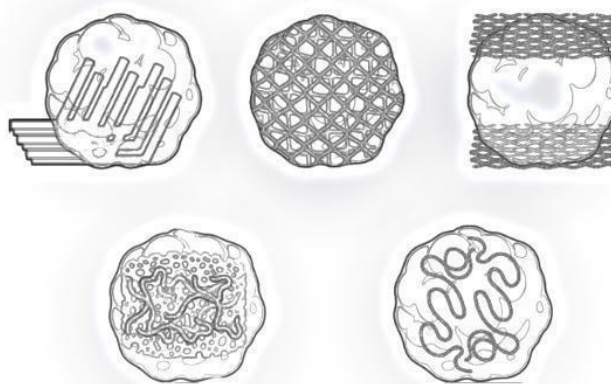
¹⁶ Smirnova et al., “Organoid Intelligence: New Frontier”

At the core of the tripartite interface, we have HBOs, which serve as functioning neural networks capable of processing information. HBOs' inherent ability to form and reorganize synaptic connections means that they can be trained, opening up new possibilities for research and application.

MEAs are used for interfacing HBOs with external systems, enabling precise delivery of electrical stimuli and recording of neuronal activity.¹⁷ Different types of MEAs, such as shank and mesh electrodes, offer specialized functionalities, with shank electrodes allowing access to deeper layers of organoids and mesh electrodes providing flexible interfaces.¹⁸ This bidirectional communication probes neuronal networks within the organoids and stimulates their development (Figure 3).

In addition to MEAs, the microfluidic platforms that house the organoids are essential to their sustained functionality. Typically housed in a petri dish, microfluidic platforms deliver a carefully balanced culture medium that supports cell growth and development.¹⁹ These platforms simulate the microvascular networks of the human brain, delivering a continuous flow of culture medium that mimics the nutrient and waste exchange found in vivo. This controlled environment ensures that the organoids remain healthy and responsive over extended periods, thus maximizing their utility in experimental setups.²⁰

Together, these technologies enable us to process and study organoid output, offering insights into their computational potential and applications in neurological research. In the following sections, we will explore multiple frameworks for OI application, tracing the evolution from neuromorphic computing to direct organ-on-chip systems.



Microelectrode Arrays for Organoids

Figure 3 Various types of MEAs for organoids (adapted from Passaro and Stice, “Electrophysiological Analysis”).

2.2 Artificial Neural Networks vs. Organoid Neural Networks

The use of OI as a new form of computation comes partly from the fact that, in learning, organoid neural networks (ONNs) are formed. These could be useful in solving a range of downstream tasks, because they solve the problem of traditional artificial neural networks (ANNs) that are inherently static systems, characterized by fixed topologies. Once an ANN is designed and trained, its structural properties—such as the number of layers, neurons, and connections—remain unchanged.

This is not just the case for simple learning paradigms—such as supervised, unsupervised, or reinforcement learning—but also for more complex subparadigms, such as continual learning (otherwise known as lifelong learning²¹), where the model can learn from new information over time, or self-supervised learning, where the model uses the data itself to generate labels.²² In all of these paradigms, the models are incapable of changing their own architecture.

This rigidity contrasts sharply with the learning in biological neural networks, such as those formed by organoids, which exhibit fluid intelligence and neuroplasticity. Unlike ANNs, the neurons in a brain or brain organoid can form new connections with other neurons, enabling not only continuous learning but also continuous adaptation of the architecture itself. This dynamic quality highlights the potential of OI, where evolving neural topologies could lead to more flexible and adaptive forms of computation.²³

¹⁷ Passaro and Stice, “Electrophysiological Analysis.”

¹⁸ Passaro and Stice, “Electrophysiological Analysis.”

¹⁹ Quintard, “Microfluidic Platform.”

²⁰ Passaro and Stice, “Electrophysiological Analysis”; Sharf et al., “Functional Neuronal Circuitry”; Quintard et al., “Microfluidic Platform”; Ballav et al., “Organoid Intelligence.”

²¹ Parisi et al., “Continual Lifelong Learning.”

²² Jaiswal et al., “Contrastive Self-Supervised Learning.”

²³ Mencattini, “Assembloid Learning.”

3 Current Applications of Organoid Intelligence

Currently, researchers are able to monitor and modulate the neural activity, effectively integrating the biological neural networks in HBOs with electronic systems.²⁴ The organoids are embedded into computational frameworks that can enable them to perform specific tasks. Neural signals from the HBOs can control virtual environments or robotic systems, enabling the study of learning and adaptive behaviors. Algorithms decode these neural patterns and optimize the interactions between organoids and their virtual or physical embodiments.²⁵

OI creates a bidirectional communication system between the HBOs and the interfaced AI system through MEAs. OIs are both “plugged in” to these AI chips and connected to the HBOs’ neural activity.²⁶ In this section, we review a (nonexhaustive) list of current OI case studies.

3.1 DishBrain Pong

The virtual and robotic embodiment of OI immerses the biological neural network within virtual or robotic environments, which allows them to interact and adapt in simulated worlds.²⁷ A prominent OI case study is the “DishBrain” device by Kagan and team at Cortical Labs,²⁸ which demonstrates how HBOs can be integrated into a simulated game environment of *Pong*.²⁹

The primary objective of DishBrain was to explore the capabilities of in vitro neural networks to perform goal-oriented tasks when provided with sensory input and feedback. Kagan and his team grew organoids on MEAs,³⁰ which recorded and stimulated the electrical activity within the neurons and involved a grid of electrodes that delivered electrical stimuli to specific regions of the neural network.³¹ DishBrain was virtually embodied into a game environment that simulated the arcade game *Pong*, where a virtual paddle controlled by the biological neural network interacts with a ball that moves back and forth across the screen as a closed-loop feedback system.³² If the virtual paddle was successful in hitting the virtual ball, a ‘positive’ feedback signal was sent to the sensory region to reinforce this behavior, and if the paddle missed the ball, a ‘negative’, less predictable feedback signal was sent. The biological neural network adapted to the feedback from the game environment and improved its performance, demonstrating learning and adaptive behavior. The research also reveals that DishBrain competes with other existing deep reinforcement learning algorithms.³³ By demonstrating that neurons can learn and adapt in goal-directed ways, DishBrain opens new avenues for brain–machine interfaces, neurocomputational models, and biological–artificial hybrid systems.

3.2 Neanderoids

Neuroscientist Alysson Muotri and his team at the University of California, San Diego, pioneered the development of Neanderthal brain organoids by reintroducing the archaic variant of the NOVA1 gene into human stem cells.³⁴ These organoids exhibit significant differences from human cortical organoids, including longer growth periods, a distinctive popcorn-like shape, and fewer cortical connections. These findings suggest that the neurological structures of Neanderthals may have influenced their cognitive abilities and social behaviors in ways that differ from those of modern humans, and this research offers a novel platform for studying human evolution and neurodevelopment. Further, Muotri and team connected Neanderthal organoids to robotic systems, allowing these brain models to interact with and explore their environment.³⁵ This fusion of biology with modern technology opens a realm of new possible research for understanding cognition, learning, and adaptation across evolutionary timescales. This work not only enhances our understanding of human brain evolution but also lays the groundwork for future studies that may uncover the genetic basis of human-specific traits and vulnerabilities.

3.3 Speech Recognition Studies

Guo and his research team at Indiana University Bloomington developed a novel hybrid system called Brainoware. In their study, published in *Nature Electronic*, Guo’s team conducted a benchmark test to evaluate Brainoware’s capabilities in speech recognition.³⁶ They used 240 audio clips of Japanese vowels, which were converted into electrical signals and processed by the HBOs. These signals were then decoded by an AI tool. Although the system initially showed low accuracy, it improved with training, eventually reaching a 78 percent accuracy rate. While this is lower than that of conventional

²⁴ Smirnova et al., “Organoid Intelligence: New Frontier.”

²⁵ Kagan et al., “In Vitro Neurons”; Muotri, “Brain Model Technology.”

²⁶ Greenberg, “Birth of Wetware.”

²⁷ Kagan et al., “In Vitro Neurons”; Smirnova et al., “Organoid Intelligence: New Frontier”; Khajehnejad et al., “Biological Neurons.”

²⁸ Kagan et al., “In Vitro Neurons.”

²⁹ Kagan et al., “In Vitro Neurons.”

³⁰ Kagan et al., “In Vitro Neurons.”

³¹ Khajehnejad et al., “Biological Neurons.”

³² Kagan et al., “In Vitro Neurons.”

³³ Khajehnejad et al., “Biological Neurons.”

³⁴ Trujillo et al., “Reintroduction.”

³⁵ Trujillo et al., “Reintroduction.”

³⁶ Cai et al., “Brain Organoid Reservoir.”

ANNs, the study is a pioneering demonstration of how organoid-based systems can learn and perform computational tasks, marking a significant step forward in the development of biocomputers.³⁷

3.4 Bioprocessors

Companies like FinalSpark and Emulate are already selling OI technologies as advanced biocomputational devices.³⁸ OI is claimed to enhance our understanding of brain function and create new forms of biocomputers that could surpass the efficiency and capabilities of traditional silicon-based systems.³⁹ As these platforms continue to evolve, they hold the potential to not only transform scientific research but also expand what is currently possible in computing and artificial intelligence, creating hybridized entities of biology and technology.⁴⁰

3.5 Internet of Organoids

FinalSpark has developed an initiative that allows real-time online monitoring of their biochips, offering a window into the live neuronal activity of organoids.⁴¹ Through their online platform, users can observe the real-time functionality of neurospheres housed within the MEAs. Individual charts displayed on the platform correspond to a single biochip, where the activity of one neurosphere is tracked. The charts provide detailed information on the electrophysiological signals detected by the electrodes in the MEA, with each signal representing the neuronal activity of the organoid in response to various stimuli. This feature represents a significant step toward integrating biological systems with digital platforms. By offering live views of the organoids' activity, FinalSpark enables researchers, students, and the public to directly witness the complexities of neuronal signaling and the potential of OI.

4 The Design Space of Organoid Intelligence

As we move from the established foundation and current applications of OI into the design space, we delve into the speculative and exploratory. This space is used to envision the future of OI, where current technological capabilities intersect with innovative concepts. This enables us to explore the potential trajectories of OI development, understanding where the hybridization of biological systems and technology could lead, without necessarily advocating for any specific evolutionary path. The importance of the design space lies in its ability to offer a creative framework to anticipate future opportunities and challenges. By engaging with the speculative, we can identify and address the current limitations of OI technologies while imagining how these challenges might be overcome.

Currently, OI technology faces several significant limitations. Some of these include the lack of vascularization in organoids, which restricts their growth and complexity. OI also has other scope and scale limitations around potential interorganoid communication and the quality of microfluidic platforms. These limitations set important boundaries on what is currently achievable, yet this design space enables us to speculate on how these limitations might be addressed in the future, through innovations in bioengineering, computational frameworks, and ethical oversight.

The future of OI looks to the design of OI systems as a guided evolution. The potential of OI can be understood through theoretical frameworks such as Friston's "free energy principle." This principle posits that the brain constantly strives to create a predictive model of the world, minimizing the gap between sensory inputs and its predictions.⁴² Applied to organoids, this suggests that their development of structures such as ocular cups indicates a "demand" for more complex sensory inputs. This self-driven complexity reveals the growing computational potential of organoids as a material substrate for intelligence.

Within this design space, we introduce the concept of scaffolding as a layered approach to developing OI systems. Scaffolding here refers to the idea that each layer of development builds on the last, creating increasingly sophisticated and capable systems. The layers of scaffolding we explore are highly speculative, extending the boundaries of current technology and envisioning how future advancements could fundamentally change what OI can achieve.

The layers we discuss include Layer 1: Organoid Array Computing, where multiple organoids work in parallel to enhance computational capacity; Layer 2: Multimodality Processing, which imagines specialized organoids designed to process different types of sensory inputs; and Layer 3: Intergenerational Memory, which speculates on organoids' potential to transmit learned behaviors or information across generations. These layers represent possible future directions in the development of OI, providing a roadmap for how these biological systems could evolve in complexity and functionality.

³⁷ Tsanni, "Human Brain Cells."

³⁸ FinalSpark, "Neuroplatform"; Emulate, "Brain-Chip."

³⁹ Smirnova et al., "Organoid Intelligence: New Frontier."

⁴⁰ Smirnova et al., "Organoid Intelligence: New Frontier"; Smirnova et al., "Organoid Intelligence: Ultimate Functionality."

⁴¹ FinalSpark, "Neuroplatform."

⁴² Friston, "Free-Energy Principle."

4.1 Current Limitations: Vascularization

Current organoid technologies face significant limitations due to HBOs' lack of vascularization. While HBOs offer a promising avenue for studying human brain development and disease, they often lack the microenvironment and vascular support necessary for sustained growth and functionality.⁴³ This deficiency results in necrotic centers due to insufficient oxygen and nutrient supply, limiting their size and complexity. Although efforts to integrate vascular structures into organoids through cocultures with vascular cells or tissue engineering have shown some promise, they have not yet achieved the authentic blood microenvironment required for proper development.⁴⁴

An alternative approach involves the engraftment of HBOs into animal hosts, such as mice, where they can develop functional vasculature and integrate with the host brain's neuronal circuits.⁴⁵ This method has shown success in creating mature and functional human brain tissues in vivo, responding to physiological stimuli and demonstrating functional synaptic connectivity. Preliminary indications are that the human neuronal tissue can not only be grafted into rodent neuronal tissue but also receives sensory input and becomes permeated by blood vessels supplying oxygen and nutrients and carting away metabolic waste.⁴⁶ However, the use of chimeras raises ethical concerns, particularly regarding the potential for these organoids to develop morally relevant qualities.⁴⁷ As the field advances, it is crucial to address these ethical issues proactively to ensure the responsible development and application of OI technologies.

Another potential solution for the significant challenge of vascularization is bioprinting. While it is commonly associated with creating skin grafts or ear transplants, recent research has employed this technique for introducing vascular structures into organoids.⁴⁸ The development of bioprinting has traversed several stages, from first bioprinting nonbiocompatible structures, to nonbiodegradable prostheses, and toward biocompatible and biodegradable structures that support tissue repair and regeneration.⁴⁹ Bioprinting uses different bioinks, such as hydrogel composed of cellulose or collagen. When cells are introduced to the microlattice scaffold, they can grow over this scaffold to turn into a 3D structure so that the bioprinted structure provides physical instructions for growth. At times, these temporary and thermoreversible supports can also be washed away, facilitating plastic development. Currently, researchers are able to print biomimic 3D structures with living cells akin to cellular printing, covering multiple materials and cell types.⁵⁰

4.2 Microfluidic Platforms

Beyond addressing vascularization challenges, we can also reconceptualize microfluidic platforms so that they are conditioned to support responsive, open-ended cell development.⁵¹ To consider the potential computational power of organoid systems beyond pure electrical stimulation, we should reposition microfluidic platforms as an architecture for non-electrical stimulation.⁵²

Here, chemical and biological computing systems recursively program each other and exhibit multiscale continuation.⁵³ A more recent concept proposed by Leroy Cronin, *chemputation*, also points to the increasing interest in metabolic design for polycomputational systems.⁵⁴ With this understanding, we can view microfluidic platforms as an additional layer of computation, working in concert with organoid structures and electrical feedback systems.

To optimize microfluidic platforms for OI, we propose to develop systems that allow for the circulation of essential nutrients and growth factors, moving beyond static culture broths. Studies show that an extended culture medium permits the development of mature cell types and cellular diversity.⁵⁵ For example, microfluidic platforms contribute to the diversity of inputs that organoids are capable of receiving, as culture conditions and duration are critical factors in neuronal maturation and functionality.⁵⁶ Neuronal plasticity can be regulated through different culture media, such as neural induction medium, DMEM, maintenance medium, Neurobasal Medium, and BrainPhys Medium.⁵⁷ Not only is neuronal activity measured and conditioned with these media, different media help optimize the exposure of fluorescent compounds to organoids and modulate downstream neuronal networks⁵⁸.

⁴³ Zhang et al., "Vascularized Organoids."

⁴⁴ Mansour et al., "In Vivo Model"; Chen et al., "Human Brain Organoids."

⁴⁵ Mansour et al., "In Vivo Model."

⁴⁶ Goldman, "Assembloid Models."

⁴⁷ Hyun et al., "Ethical Issues."

⁴⁸ Kengla et al., "Bioprinting of Organoids"; Wang et al., "Application of Bioprinting"; Ren et al., "Developments and Opportunities."

⁴⁹ Wang et al., "Application of Bioprinting."

⁵⁰ Wang et al., "Application of Bioprinting"; Skylar-Scott et al., "Orthogonal Differentiation."

⁵¹ Quadrato et al., "Cell Diversity"; Cogoni et al., "ISiCell."

⁵² Smirnova et al., "Organoid Intelligence: New Frontier"

⁵³ Bongard and Levin, "Biological Systems."

⁵⁴ Cronin, "Chemputer and Chemputation"; Sha, "Metabolic Approach."

⁵⁵ Quadrato et al., "Cell Diversity."

⁵⁶ Osaki et al., "Complex Activity."

⁵⁷ Zabolocki et al., "BrainPhys Neuronal Medium."

⁵⁸ Zabolocki et al., "BrainPhys Neuronal Medium" 2020; Osaki et al., "Complex Activity." 2024

The Chemputer proposed by Cronin is a speculative universal chemical synthesis machine that automates the precise control of chemical reactions, using programmable hardware to execute complex synthesis pathways.⁵⁹ This polycomputational approach suggests that if liquids can compute, the chemical microenvironments within microfluidic platforms contribute to the overall computational capacity of the system by shaping the conditions under which biological processes occur. Similar to how the Chemputer uses precise control of reagents and conditions to guide chemical reactions and achieve desired outputs, microfluidic platforms in OI systems could dynamically regulate chemical gradients, nutrient delivery, and waste removal. This controlled environment enables the fine-tuning of organoid development and neural activity, effectively guiding the biological processes that underlie computational tasks. In this sense, the liquid environment becomes an integral part of the system's computational framework, enabling the organoids to perform more complex functions by optimizing the conditions for their growth and interaction.

Another branch of organoid engineering research points to hormonal alteration and dopamine stimulation as additional inputs that can influence organoid growth and development.⁶⁰ The stimulation of dopaminergic neurons could entrain and modify the activity patterns of neurons in other regions of the assembloid and induce long-lasting morphological changes.⁶¹

Through these systems, we may also explore a range of substrate materials and configurations to promote open-ended cell development. Further, we might want to design microfluidic platforms that facilitate bidirectional communication between electric signals and chemical culture components. Recursive chemical-to-electrical artificialization enhances the organoids' ability to adapt, self-programming the development of computational capabilities. This fluid architecture is necessary to support the neuroplasticity and adaptability of ONNs.

4.3 Interorganoid Communication: From Organoids to Assembloids

Organoids are often cultured to grow with closely monitored factors to ensure experiment reproducibility and reliability of electrical recording results.⁶² However, most experiments are conducted on a specific type of organoid, lacking the ability to mimic fully matured adult brains that have interregional and intercellular interactions.⁶³ In relation to organoids' potential for heterogeneous development, Sergiu Pașca, director of Stanford's Brain Organogenesis Program, attributes this phenomenon to the brain's inherent capabilities to self-organize "with its own assembly instructions."⁶⁴

An emerging engineering approach is the combination of different region-specific neural organoids into fusion or assemblies to recapitulate the interaction between brain regions. Fused organoids can mimic neural migration, projection, or functional neural circuits between brain regions.⁶⁵ Meanwhile, researchers have also started coculturing differently patterned organoids or combining neural organoids with nonneural tissues to model cell migration and connectivity.⁶⁶ For example, Shi and team generated vascularized human cortical organoids (vOrganoids) by coculturing human embryonic stem cells or human-induced pluripotent stem cells with human umbilical vein endothelial cells in vitro.⁶⁷ There is also an assembly of blood vessels and brain cells and a trio of cerebral cortex, spinal cord, and muscle organoids demonstrating orchestration.⁶⁸ Networking assembloids have shown a sign of neuroplasticity through "short-term potentiation," supporting the fluidity of cognitive architectures.⁶⁹

Beyond fused and assembled organoids, HBOs can also communicate with each other by forming connections through axons, the long, thread-like extensions of neurons that transmit electrical signals.⁷⁰ Current technology enables researchers to cultivate reciprocal axon bundles between organoids using specialized silicon elastomer microdevices that provide a microchannel to guide the growth of these connections.⁷¹ These connections have been shown to transmit electrical impulses from one organoid to another, demonstrating a form of communication without external molecular instructions. This research reveals that spatial instructions alone (by design of the microdevice) can sufficiently direct organoid development and regeneration (Figure 4). The design of the microdevice, for example, the number of units and dimensions of its channels, and the choice of biocompatible materials provide physical instructions by determining and structuring certain mechanical forces such as compression, pressure, etc.⁷²

⁵⁹ Cronin, "Chemputer and Chemputation."

⁶⁰ Reumann et al., "In Vitro Modeling."

⁶¹ Reumann et al., "In Vitro Modeling."

⁶² Chung et al., "Electrophysiological Recording Platforms."

⁶³ Makrygianni and Chrousos, "From Brain Organoids."

⁶⁴ Goldman, "Assembloid Models."

⁶⁵ Bagley et al., "Fused Cerebral Organoids" 2017; Suong et al., "Design of Neural Organoids." 2024

⁶⁶ Levy and Pașca, "What Have Organoids."

⁶⁷ Shi et al., "Vascularized Human Cortical Organoids."

⁶⁸ Bagley et al., "Fused Cerebral Organoids" 2017; Birey et al., "Human Forebrain Spheroids." 2017

⁶⁹ Osaki et al., "Complex Activity."

⁷⁰ Kirihaara et al., "Model of a Cerebral Tract."

⁷¹ Kirihaara et al., "Model of a Cerebral Tract."

⁷² Kirihaara et al., "Model of a Cerebral Tract."

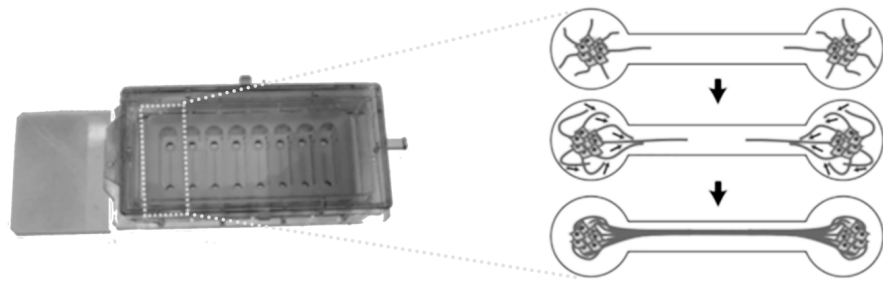


Figure 4 Axon fascicle formation from microdevice physical cues (adapted from Kiriha et al., “Model of a Cerebral Tract.”)

When two organoids are connected by an axon bundle, they perform a ‘handshake’ that leads to synchronized bursts of electrical activity recorded between two organoids.⁷³ There are existing experiments that support the claim that these interorganoid axonal connections not only correlate to higher short-term plasticity in the neuronal network but also facilitate the development of a higher complexity of signals between connected organoids.⁷⁴

At present, the connections are limited to direct electrical signaling without the intricate synaptic networks found in a fully developed brain. Speculatively, advancing this technology could involve creating more complex microenvironments that promote the formation of more sophisticated neural circuits, including synapses and potentially even chemical signaling pathways. By incorporating factors that encourage the development of these connections, such as growth factors, and using more advanced bioengineering techniques, we could potentially create organoid systems that mimic the complexity of real brain networks and develop robust communication for mutual learning. This could eventually lead to the creation of interconnected organoid networks capable of more advanced, coordinated activities, offering profound insights into brain function and the mechanisms of neurological diseases.

5 Scaffolding for Organoid Intelligence

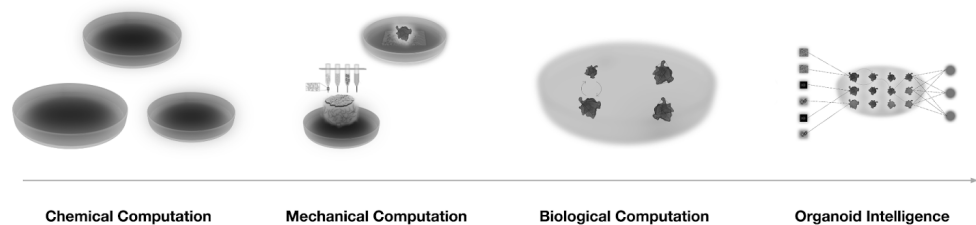


Figure 5 Scaffolding for OI: Chemical computation: designing protocols for culture media and chemical stimuli to influence organoid development, reproducibility, and function; mechanical computation: designing mechanical properties of tissues, biomaterials, and interfaces to provide spatial conditioning and physical cues for organoid growth, e.g. materials, electrophysiological factors, solubility, bioprinting scaffold designs, dimensional parameters, MEAs, and interfacing techniques; biological computation: designing arrays of cellular growth, tissue morphogenesis, maturation, and assembly as well as developing various organoid types; OI: integrates chemical, mechanical, biological, and electrophysiological signal processing to create a comprehensive framework for OI. Figure by Jenn Leung.

Scaffolding for OI relies on a range of dependencies, including microfluidics for nutrient delivery, electric pulses for stimulation, and physical hardware support such as MEAs. However, once these dependencies are in place, HBOs begin to form their own functional neural networks, effectively becoming a form of biological hardware capable of supporting higher-order computational tasks. This continuum of scales—from chemical microenvironments, to mechanical spatial conditioning, tissue

⁷³ Osaki et al., “Complex Activity.”

⁷⁴ Osaki et al., “Complex Activity.”

scaffolding, and finally, OI—forms the basis of a polycomputational system, where different layers of computation, both biological and artificial, interact to create a complex and adaptable system (Figure 5).⁷⁵ Fluid intelligence and neuroplasticity is scaffolding for an OI that has evolving topologies.⁷⁶

To further explore the potential of OI, additional experiments would help refine the encoding and decoding of spatiotemporal information within these neural networks. This can include exploring machine learning frameworks such as reinforcement learning (RL) and reservoir computing (RC), which may help unlock the next phases of OI. By scaffolding these experimental designs, we move to examine how organoids might evolve from simple neural assemblies to systems potentially capable of multimodal processing, intergenerational memory transfer, and polycomputational functionality—where multiple computational paradigms operate simultaneously within a unified system.

5.1 Layer 1: Organoid Array Computing

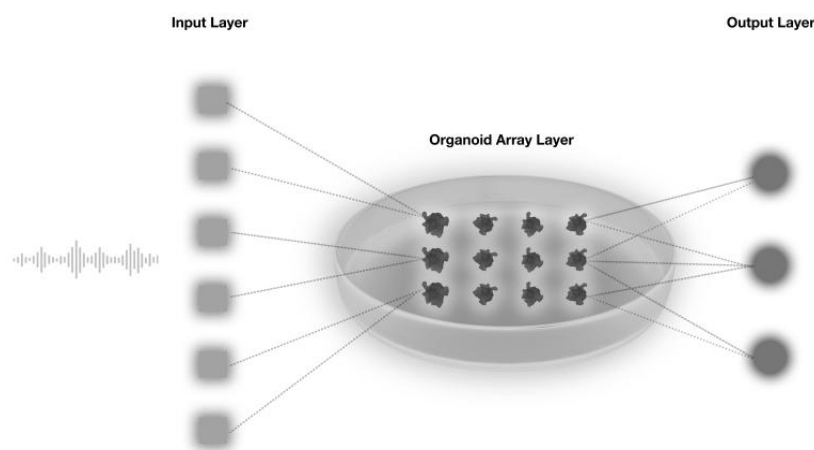


Figure 6 Design Layer 1: Organoid Array Computing. Input layer: the input layer converts information (image pattern, audio clips, time series, etc.) into various spatiotemporal sequences of electrical stimulation pulses; organoid array layer: the organoid array receives the input electrical stimulation and maps it to a high-dimensional computational space as the ONN; decoder layer: neural activities are fed into decoding functions such as linear regression or logistic regression to form an output layer for classification, recognition, and prediction. Figure by Jenn Leung.

Organoid array computing presents a biocomputing architecture that employs neural assemblies as physical computational systems. This approach aligns with the concept of RC, an energy-efficient method that utilizes an untrained reservoir and a linearly trained simple classifier.⁷⁷

Our proposed design layer envisions a three-dimensional assembly of HBOs functioning as a nested ONN, serving as the reservoir in an RC framework for speech recognition tasks. Using a biological neural network as a form of reservoir in an RC framework, we could evaluate the feasibility of using a multi-organoid array as an RC system for speech recognition tasks.⁷⁸

Building on recent advancements, including Brainware’s interface’s success in vowel identification and Cortical Labs’ exploration of multi-organoid arrays for memory storage, we devised potential future proof-of-concept for scaffolding for an expanding OI design space.⁷⁹ The set-up would include an array of HBOs, derived from induced pluripotent stem cells. This organoid array is housed in a custom-designed microfluidic platform that enables nutrient perfusion and potential interorganoid chemical signaling. Each organoid in the array is interfaced with a high-density microelectrode array (HD-MEA) for stimulation and recording. The entire system is maintained in an environmental control system to ensure optimal conditions for organoid health and function. A signal processing unit converts audio inputs into electrical stimuli, while a machine learning interface implements a linear classifier for decoding organoid responses (Figure 6).

We can consider the input layer as a layer that converts information (image pattern, audio clips, time series data) into various spatiotemporal sequences of electrical stimulation pulses that can be sent to the organoid.⁸⁰ It is shown in the DishBrain experiment that inputting electrophysiological input through eight stimulation electrodes with rate coding along with place coding electrical pulses to communicate

⁷⁵ Zhang et al., “Translational Organoid Technology.”

⁷⁶ Bakkum et al., “Activity-Dependent Plasticity.”

⁷⁷ Glover et al., “Reservoir Computing.”

⁷⁸ Cai et al., “Brain Organoid Reservoir.”

⁷⁹ Cai et al., “Brain Organoid Reservoir.”

⁸⁰ Khajehnejad et al., “Biological Neurons.”

bounded two-dimensional data is comparable and outperforms pixel-based information input to RL algorithms (Figure 7).⁸¹ We can attempt to extend this experiment to cater for other inputs.

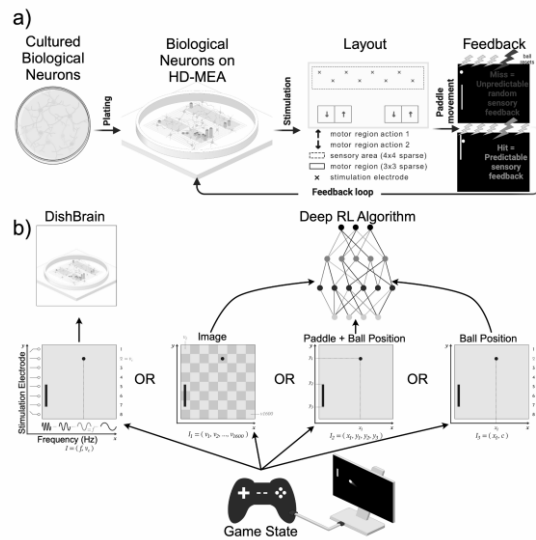


Figure 1: DishBrain system and Various input designs to RL algorithms. a) DishBrain feedback loop setup and Electrode configuration and predefined sensory and motor regions. Figures adapted and modified from (Kagan et al., 2022). b) Schematic comparing the information input routes in the DishBrain system (left) and the three implementations of the deep RL algorithms (right). In each design, the input information to the computing module (deep RL algorithms or DishBrain) is denoted by a vector I .

Figure 7 DishBrain system and various input designs to RL algorithms.⁸²

In the organoid array computing layer, dozens of miniature HBOs are housed in the ONN, each serving as a node, each interfaced with an MEA and connected through a multichambered microfluidic device. As electrical stimuli (converted from audio inputs) are applied to the organoids via HD-MEAs, the complex neural networks within each organoid transform these inputs into a higher-dimensional representation (Figure 6). This is a fundamental property of RC, where the reservoir (in this case, the organoid array) projects the input into a high-dimensional space.

In the output/decoder layer, neural activities representing the state of the ONN are recorded by an MEA system and fed into decoding functions. This makes the output readable for downstream tasks, forming an output layer for classification, recognition, prediction, and other applications (Figure 6).⁸³

The application of RC to OI expands the dimensions of biocomputing capabilities in the design space for experiments. As RC is a substrate-independent framework, it enables the expansive integration of various components into OI systems, including organoid arrays, microfluidic platforms, HD-MEAs, environmental control systems, signal processing units, and machine learning interfaces. The paper “Assembloid Learning” proposes that “personalized models” of neural assemblies could be developed in the near future for brain care and treatment optimization.⁸⁴ This possibility arises from the adaptive nature of neural assemblies, which can learn and respond to chemical and electrical stimuli that induce plastic changes, especially when derived from an individual’s own cells.

The implementation of a multi-organoid RC framework in OI has several important implications for biocomputing. First, it enables assembloid learning, where multiple organoids can coordinate efforts, potentially mimicking the distributed processing of the human brain.⁸⁵ Second, the parallel processing capabilities of multiple organoids in an array could significantly increase computational capacity, enabling more complex task pooling and information processing, scaffolding for a task-pooling intelligence among organoids in the array. Additionally, as organoids have shown the ability to develop regional specialization similar to the human brain, this framework could lead to more sophisticated modeling of brain functions, leading to its expanding capabilities to process multimodal inputs.⁸⁶ Finally, the ability to process simulated environments, as demonstrated in experiments such as the “DishBrain” Pong game, opens up possibilities for creating virtual environments encoded as spatiotemporal electrophysiological activity.⁸⁷ These developments collectively suggest that OI has the potential to scaffold for multimodal and multilayered environments.

⁸¹ Khajehnejad et al., “Biological Neurons.”

⁸² Khajehnejad et al., “Biological Neurons.”

⁸³ Cai et al., “Brain Organoid Reservoir.”

⁸⁴ Mencattini, “Assembloid Learning.”

⁸⁵ Mencattini, “Assembloid Learning.”

⁸⁶ Sun et al., “Translational Potential.”

⁸⁷ Kagan et al., “In Vitro Neurons.”

5.2 Layer 2: Scaffolding for Multimodality

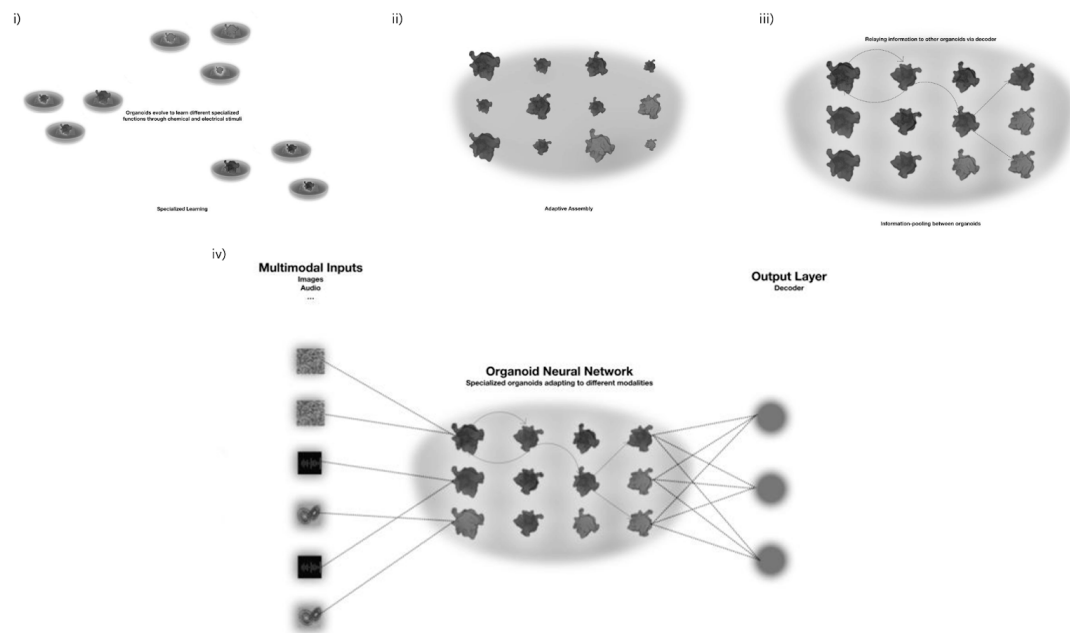


Figure 8 Design Layer 2: Multimodality Processing. (i) Organoids evolve to learn different specialized functions through chemical and electrical stimuli; (ii) adaptive assembly among HBOs; (iii) relaying information to other organoids via decoder; (iv) Design Layer 2: Scaffolding for multimodality. Figure by Jenn Leung.

Building on the foundation of organoid array computing, we can envision a more advanced design layer that leverages the potential of multiple organoids to process multimodal inputs. Recent advancements in brain organoid research, such as region-specific HBOs, vascularized organoids, and assembloids (which combine different organoid types) provide a promising platform for this next step in OI.⁸⁸ By cultivating specialized structures that mimic the sensory regions of the brain, we can envision a system where different organoids within an array are designed to handle distinct types of sensory information (Figure 8).

For example, some HBOs could be guided to specialize in processing tailored inputs. This specialization could be achieved by exposing the organoids to specific inputs or stimulation techniques, such as electrical signals via MEAs, optogenetic manipulation, or chemical stimulation. With this approach, each organoid would act as a specialized processing unit, much like how the human brain allocates distinct regions to handle sensory inputs like sight, sound, and touch.⁸⁹

If specialized organoids are developed, they could be integrated into an interconnected system where multimodal inputs are processed simultaneously. By employing different stimulation techniques to cater to each organoid's specialized function, the entire array would “collaborate” to interpret complex, multisensory inputs. This mirrors the brain's ability to integrate information from multiple sensory modalities to form a cohesive understanding of the environment.

Taking assembloid learning as a key path dependency from Layer 1, this layer suggests that HBOs could be cultivated to mimic the brain's inherent ability to process multimodal inputs. By utilizing one RC framework to handle this data, the organoid array would draw parallels between, for example, visual pixels, audio patterns, and touch-based stimuli. Such a system could extend the current capacities of OI, offering a design that not only processes individual data streams but also integrates them into a unified output. This paves the way for future experiments that explore how organoids could be trained to develop specialized functions adaptable to more complex, multimodal computing tasks.

⁸⁸ Schmidt, “Rise of the Assembloid” 2021; Mansour et al., “In Vivo Model” 2018; Susaimanickam et al., “Region Specific Brain Organoids.” 2022

⁸⁹ Ackerman, *Discovering the Brain*.

5.3 Layer 3: Intergenerational Memory

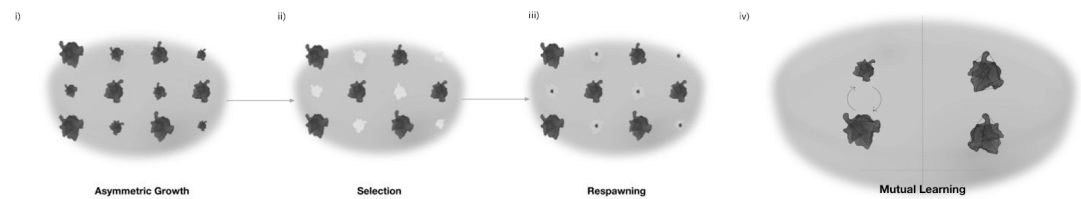


Figure 9 Design Layer 3: Intergenerational Memory. (i) Asymmetric growth with varying responses to different stimuli; (ii) HBOs go through a process of selection and survival throughout training process; (iii) respawning and optimizing HBOs; iv) mutual learning, where HBOs take into account the decoded response of other organoids during training, thereby growing new structures.

The speculative design space for intergenerational organoid communication is grounded in the emerging capabilities of HBOs to store and transmit complex information.⁹⁰ As organoids are increasingly capable of demonstrating aspects of memory storage and differentiation,⁹¹ there is a potential for creating systems where these “memories” or learned behaviors could be passed from one organoid to another, or even across generations of organoids.

Intergenerational organoid communication could involve the transfer of encoded information from one organoid to another, simulating the biological inheritance of cognitive abilities. This could be achieved through advanced bioengineering platforms that enable the selective transfer of neural patterns, potentially akin to how synaptic plasticity underpins memory in the human brain (Figure 9).⁹²

The potential applications of intergenerational memory in organoids could reconstruct how we understand and utilize biological intelligence. An organoid network could be designed where early-generation organoids undergo specific training or learning tasks, with subsequent generations inheriting this trained state, thus reducing the time and energy required for each new generation to achieve the same level of functionality. This would not only enhance the efficiency of organoid-based systems but could also lead to breakthroughs in understanding how memory and learning are encoded biologically. Such advancements might open new avenues for developing biocomputers that continuously evolve and adapt over time, integrating past experiences into future decision-making processes.

6 Implications

Advancements in organoid technologies have the potential to vastly benefit humanity. This research offers insights into human biology, disease mechanisms, and cognitive processes, and it holds promise for revolutionary medical treatments and drug development. Given these benefits, it would be misguided, and harmful, to prematurely halt or excessively restrict this field of study.

However, as we push the boundaries of what’s possible with organoid development, we are presented with ethical challenges that require thoughtful ethical analysis. The potential emergence of morally relevant qualities, such as consciousness or advanced cognitive abilities, require due ethical consideration. Continued ethical assessment will be essential as this technology advances, requiring the establishment of clear criteria to identify morally relevant capacities in HBOs and guidelines for responding appropriately to their emergence. By doing so, we can advance this promising field responsibly, maximizing its benefits to humanity while mitigating ethical risks.

A recent letter titled “A Response to Claims of Emergent Intelligence and Sentience in a Dish” underscores the importance of cautious and precise language when describing the capabilities of neural systems.⁹³ The authors criticize the premature attribution of terms such as “sentience” and “intelligence” to neurons in a dish, emphasizing that such claims lack sufficient evidence and risk creating confusion around the ethical implications of this research. We must be vigilant in how we communicate OI’s capabilities, ensuring that we do not oversell findings or trigger concerns before they are warranted.

7 Conclusion

This paper considers the expansive design space of OI, providing a comprehensive application framework of OI. Our research has aimed to provide a taxonomy of design possibilities, considering various dimensions and parameters of inputs and outputs. These include alternative configurations of organoid assemblies, mechanical devices, microfluidics, bioprinting techniques, and culture media, all of which contribute to the complex ecosystem of OI. Although OI is still in its infancy, researchers could follow the design framework to devise experiments that are adaptable to the growing dimensions of

⁹⁰ Smirnova et al., “Organoid Intelligence: New Frontier.”

⁹¹ Cai et al., “Brain Organoid Reservoir” 2023; Smirnova et al., “Organoid Intelligence: New Frontier.” 2023

⁹² Kennedy, “Synaptic Signaling.”

⁹³ Balci et al., “Response to Claims.”

neural assemblies, facilitating plastic changes during the development of research. This plasticity enables the creation of personalized models for OI, which could revolutionize our approach to understanding and manipulating these intricate biological systems.

As researchers face the challenge of isolating and monitoring specific inputs, outputs, and developmental stages to gain a deeper understanding of how HBOs function, the design space points to the nested complex system as a growing and self-patterning hybrid entity.

While the field of OI often emphasizes the evolved rather than designed nature of intelligence, it's important to note the growing body of research in evolutionary computation and artificial life. These fields reveal the potential for mechanical and refractive computing in various substrates, including granular matter.

Ultimately, the future of OI lies in the integration of multiple computational paradigms. Chemical computing, biological computing, and scaffolding systems together create a nested, multiscale system for OI. With the integration of diverse computational approaches such as RC with living neural networks, we are starting to understand forms of intelligence that more closely mirror the complexity of biological brains. To imagine the future of OI, we must also imagine growing cognitive assemblies with the capacity for physical evolution. Unlike traditional computing systems with fixed hardware architectures, organoid arrays can grow, adapt, and reorganize their physical structure in response to computational demands.

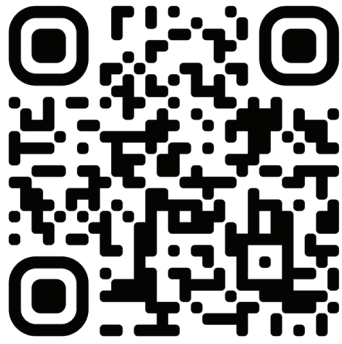
Bibliography

- Ackerman, Sandra, and National Academy Institute of Medicine. *Discovering the Brain*. National Academies Press, 1992.
- Bagley, Joshua A., Daniel Reumann, Shan Bian, Julie Lévi-Strauss, and Juergen A. Knoblich. “Fused Cerebral Organoids Model Interactions between Brain Regions.” *Nature Methods* 14, no. 7 (July 2017): 743–51. <https://doi.org/10.1038/nmeth.4304>.
- Bakkum, Douglas J., Zenas C. Chao, and Steve M. Potter. “Long-Term Activity-Dependent Plasticity of Action Potential Propagation Delay and Amplitude in Cortical Networks.” *PLoS One* 3, no. 5 (May 2008): e2088. <https://doi.org/10.1371/journal.pone.0002088>.
- Balci, Fuat, Suliann Ben Hamed, Thomas Boraud, et al. “A Response to Claims of Emergent Intelligence and Sentience in a Dish.” *Neuron* 111, no. 5 (March 2023): 604–5. <https://doi.org/10.1016/j.neuron.2023.02.009>.
- Baldassari, Simona, Ilaria Musante, Michele Iacomino, Federico Zara, Vincenzo Salpietro, and Paolo Scudieri. “Brain Organoids as Model Systems for Genetic Neurodevelopmental Disorders.” *Frontiers in Cell and Developmental Biology* 8 (October 2020): 590119. <https://doi.org/10.3389/fcell.2020.590119>.
- Ballav, Sangeeta, Amit Ranjan, Shubhayan Sur, and Soumya Basu. “Organoid Intelligence: Bridging Artificial Intelligence for Biological Computing and Neurological Insights.” In *Technologies in Cell Culture – A Journey From Basics to Advanced Applications*, edited by Soumya Basu, Amit Ranjan, and Shubhayan Sur. IntechOpen, June 2024. <https://doi.org/10.5772/intechopen.114304>.
- Beaulieu, Shawn, Piper Welch, Atoosa Parsa, Corey O’Hern, Rebecca Kramer-Bottligio, and Josh Bongard. “Refractive Computation: Parallelizing Logic Gates across Driving Frequencies in a Mechanical Polycomputer.” In *ALIFE ’24: Proceedings of the 2024 Artificial Life Conference*, edited by Andrés Faiña, Sebastian Risi, Eric Medvet, et al. MIT Press, 2024. https://doi.org/10.1162/isal_a_00807.
- Birey, Fikri, Jimena Andersen, Christopher D. Makinson, et al. “Assembly of Functionally Integrated Human Forebrain Spheroids.” *Nature* 545, no. 7652 (May 2017): 54–59. <https://doi.org/10.1038/nature22330>.
- Blakemore, Erin. “Human-Pig Hybrid Created in the Lab—Here Are the Facts.” *National Geographic*, January 26, 2017. <https://www.nationalgeographic.com/science/article/human-pig-hybrid-embryo-chimera-organs-health-science>.
- Bongard, Joshua, and Michael Levin. “Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior.” *Frontiers in Ecology and Evolution* 9 (March 2021): 650726. <https://doi.org/10.3389/fevo.2021.650726>.
- . “There’s Plenty of Room Right Here: Biological Systems as Evolved, Overloaded, Multi-Scale Machines.” *Biomimetics* 8, no. 1 (March 2023): 110. <https://doi.org/10.3390/biomimetics8010110>.
- Cai, Hongwei, Zheng Ao, Chunhui Tian, et al. “Brain Organoid Reservoir Computing for Artificial Intelligence.” *Nature Electronics* 6, no. 12 (December 2023): 1032–39. <https://doi.org/10.1038/s41928-023-01069-w>.
- Chen, H. Isaac, John A. Wolf, Rachel Blue, Mingyan Maggie Song, Jonathan D. Moreno, Guo-Li Ming, and Hongjun Song. “Transplantation of Human Brain Organoids: Revisiting the Science and Ethics of Brain Chimeras.” *Cell Stem Cell* 25, no. 4 (October 2019): 462–72. <https://doi.org/10.1016/j.stem.2019.09.002>.
- Chung, Won Gi, Enji Kim, Hayoung Song, et al. “Recent Advances in Electrophysiological Recording Platforms for Brain and Heart Organoids.” *Advanced NanoBioMed Research* 2, no. 1 (December 2022): 2200081. <https://doi.org/10.1002/anbr.202200081>.

- Cogoni, Florian, David Bernard, Roxana Kazhen, Salvatore Valitutti, Valérie Lobjois, and Sylvain Cussat-Blanc. "ISiCell: A Participatory Methodology and Platform for Collaborative Agent-Based Modeling in Cell Biology." In *ALIFE '24: Proceedings of the 2024 Artificial Life Conference*, edited by Andrés Faiña, Sebastian Risi, Eric Medvet, et al. MIT Press, 2024.
- Cronin, Leroy. "The Chemputer and Chemputation: A Universal Chemical Compound Synthesis Machine." Preprint, arXiv, August 17 2024. <https://doi.org/10.48550/arXiv.2408.09171>.
- Emulate. "Brain-Chip." Accessed May 2021. <https://emulatebio.com/brain-chip/>.
- Fernandes, Tiago G. "Organoids as Complex (Bio)Systems." *Frontiers in Cell and Developmental Biology* 11 (August 2023): 1268540. <https://doi.org/10.3389/fcell.2023.1268540>.
- Final Spark. "Neuroplatform – FinalSpark." Accessed April 2024. <https://finalspark.com/neuroplatform/>.
- Friston, Karl. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11, no. 2 (February 2010): 127–38. <https://doi.org/10.1038/nrn2787>.
- Gabriel, Elke, Walid Albanna, Giovanni Pasquini, et al. "Human Brain Organoids Assemble Functionally Integrated Bilateral Optic Vesicles." *Cell Stem Cell* 28, no. 10 (October 2021): 1740–1757.e8. <https://doi.org/10.1016/j.stem.2021.07.010>.
- Glover, Tom, Evgeny Osipov, and Stefano Nichele. "On When Is Reservoir Computing with Cellular Automata Beneficial?" Preprint, arXiv, June 13, 2024. <https://doi.org/10.48550/arXiv.2407.09501>.
- Goldman, Bruce. "Here Come the Assembloids." *Stanford Medicine Magazine*, July 27, 2022. <https://stanmed.stanford.edu/brain-tissue-assembloids-expand-brain-understanding/>.
- Greenberg, Alissa. "The Birth of Wetware." *proto.life*, May 2018. <https://proto.life/2018/05/the-birth-of-wetware/>.
- Hyun, Insoo, J. C. Scharf-Deering, and Jeantine E. Lunshof. "Ethical Issues Related to Brain Organoid Research." *Brain Research* 1732 (April 2020): 146653. <https://doi.org/10.1016/j.brainres.2020.146653>.
- Jaiswal, Ashish, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. "A Survey on Contrastive Self-Supervised Learning." *Technologies* 9, no. 1 (2021): 2. <https://doi.org/10.3390/technologies9010002>.
- Kagan, Brett J., Andy C. Kitchen, Nhi T. Tran, et al. "In Vitro Neurons Learn and Exhibit Sentience When Embodied in a Simulated Game-World." *Neuron* 110, no. 23 (December 2022): 3952–3969.e8. <https://doi.org/10.1016/j.neuron.2022.09.001>.
- Kengla, Carlos, Anthony Atala, and Sang Jin Lee. "Bioprinting of Organoids." In *Essentials of 3D Biofabrication and Translation*, edited by Anthony Atala and James J. Yoo. Elsevier, 2015.
- Kennedy, Mary B. "Synaptic Signaling in Learning and Memory." *Cold Spring Harbor Perspectives in Biology* 8, no. 2 (December 2013): a016824. <https://doi.org/10.1101/cshperspect.a016824>.
- Khajehnejad, Moein, Forough Habibollahi, Aswin Paul, Adeel Razi, and Brett J. Kagan. "Biological Neurons Compete with Deep Reinforcement Learning in Sample Efficiency in a Simulated Gameworld." Preprint, arXiv, May 27, 2024. <https://doi.org/10.48550/arXiv.2405.16946>.
- Kirihara, Takaaki, Zhongyue Luo, Siu Yu A. Chow, et al. "A Human Induced Pluripotent Stem Cell-Derived Tissue Model of a Cerebral Tract Connecting Two Cortical Regions." *iScience* 14 (April 2019): 301–11. <https://doi.org/10.1016/j.isci.2019.03.012>.
- Kruszelnicki, Karl S. "Mouse with Human Ear." *ABC Science*, June 2, 2006. <https://www.abc.net.au/science/articles/2006/06/02/1644154.htm>.
- Levy, Rebecca J., and Sergiu P. Pașca. "What Have Organoids and Assembloids Taught Us About the Pathophysiology of Neuropsychiatric Disorders?" *Biological Psychiatry* 93, no. 7 (April 2023): 632–41.

- Makrygianni, Evanthia A., and George P. Chrousos. "From Brain Organoids to Networking Assembloids: Implications for Neuroendocrinology and Stress Medicine." *Frontiers in Physiology* 12 (June 2021): 621970. <https://doi.org/10.3389/fphys.2021.621970>.
- Mansour, Abed Alfatah, J. Tiago Gonçalves, Cooper W. Bloyd, et al. "An In Vivo Model of Functional and Vascularized Human Brain Organoids." *Nature Biotechnology* 36, no. 5 (June 2018): 432–41. <https://doi.org/10.1038/nbt.4127>.
- Mencattini, Arianna, Elena Daprati, David Della-Morte, Fiorella Guadagni, Federica Sangiuolo, and Eugenio Martinelli. "Assembloid Learning: Opportunities and Challenges for Personalized Approaches to Brain Functioning in Health and Disease." *Frontiers in Artificial Intelligence* 7 (April 2024): 1385871. <https://doi.org/10.3389/frai.2024.1385871>.
- Muotri, Alysson R. "Brain Model Technology and Its Implications." *Cambridge Quarterly of Healthcare Ethics* 32, no. 4 (2023): 597–601. <https://doi.org/10.1017/S096318012300018X>.
- Osaki, Tatsuya, Tomoya Duenki, Siu Yu A. Chow, et al. "Complex Activity and Short-Term Plasticity of Human Cerebral Organoids Reciprocally Connected with Axons." *Nature Communications* 15, no. 1 (April 2024): 2945. <https://doi.org/10.1038/s41467-024-46787-7>.
- Parisi, German I., Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. "Continual Lifelong Learning with Neural Networks: A Review." *Neural Networks* 113 (2019): 54–71. <https://doi.org/10.1016/j.neunet.2019.01.012>.
- Passaro, Austin P., and Steven L. Stice. "Electrophysiological Analysis of Brain Organoids: Current Approaches and Advancements." *Frontiers in Neuroscience* 14 (2020): 622137. <https://doi.org/10.3389/fnins.2020.622137>.
- Quadrato, Giorgia, Tuan Nguyen, Evan Z. Macosko, et al. "Cell Diversity and Network Dynamics in Photosensitive Human Brain Organoids." *Nature* 545, no. 7652 (May 2017): 48–53. <https://doi.org/10.1038/nature22047>.
- Quintard, Clément, Emily Tubbs, Gustav Jonsson, et al. "A Microfluidic Platform Integrating Functional Vascularized Organoids-on-Chip." *Nature Communications* 15, no. 1 (February 2024): 1452. <https://doi.org/10.1038/s41467-024-45710-4>.
- Ren, Ya, Xue Yang, Zhengjiang Ma, et al. "Developments and Opportunities for 3D Bioprinted Organoids." *International Journal of Bioprinting* 7, no. 3 (June 2021): 364. <http://doi.org/10.18063/ijb.v7i3.364>.
- Reumann, Daniel, Christian Krauditsch, Maria Novatchkova, et al. "In Vitro Modeling of the Human Dopaminergic System Using Spatially Arranged Ventral Midbrain-Striatum-Cortex Assembloids." *Nature Methods* 20, no. 12 (December 2023): 2034–47. <https://doi.org/10.1038/s41592-023-02080-x>.
- Richards, Blake A., and Timothy P. Lillicrap. "The Brain-Computer Metaphor Debate Is Useless: A Matter of Semantics." *Frontiers in Computational Science* 4 (February 2022): 810358. <https://doi.org/10.3389/fcomp.2022.810358>.
- Sato, Toshiro, Robert G. Vries, Hugo J. Snippert, et al. "Single Lgr5 Stem Cells Build Crypt-Villus Structures In Vitro without a Mesenchymal Niche." *Nature* 459, no. 7244 (May 2009): 262–65. <https://doi.org/10.1038/nature07935>.
- Schmidt, Charlie. "The Rise of the Assembloid." *Nature* 597, no. 7878 (September 2021): S22–23. <https://doi.org/10.1038/d41586-021-02628-x>.
- Sha, Xin Wei. "Metabolic Approach to Designing Space." In *The Space of Technicity, Theorising Social, Technical and Environmental Entanglements*, edited by Robert Gorny, Stavros Kousoulas, Dulmini Perera, and Andrej Radman. TU Delft OPEN Publishing and Jap Sam Books, 2024.
- Sharf, Tal, Tjitse van der Molen, Stella M. K. Glasauer, et al. "Functional Neuronal Circuitry and Oscillatory Dynamics in Human Brain Organoids." *Nature Communications* 13, no. 1 (July 2022): 4403. <https://doi.org/10.1038/s41467-022-32115-4>.

- Shi, Yingchao, Le Sun, Mengdi Wang, et al. "Vascularized Human Cortical Organoids (vOrganoids) Model Cortical Development in Vivo." *PLoS Biology* 18, no. 5 (May 2020): e3000705. <https://doi.org/10.1371/journal.pbio.3000705>.
- Skylar-Scott, Mark A., Jeremy Y. Huang, Aric Lu, et al. "An Orthogonal Differentiation Platform for Genomically Programming Stem Cells, Organoids, and Bioprinted Tissues." Preprint, bioRxiv, July 2020. <http://dx.doi.org/10.1101/2020.07.11.198671>.
- Smirnova, Lena, Brian S. Caffo, David H. Gracias, et al. "Organoid Intelligence (OI): The New Frontier in Biocomputing and Intelligence-in-a-Dish." *Frontiers in Science* 1 (February 2023): 1017235. <https://doi.org/10.3389/fsci.2023.1017235>.
- Smirnova, Lena, Itzy E. Morales Pantoja, and Thomas Hartung. "Organoid Intelligence (OI) - The Ultimate Functionality of a Brain Microphysiological System." *ALTEX* 40, no. 2 (2023): 191–203. <https://doi.org/10.14573/altex.2303261>.
- Sun, Alfred X., Huck-Hui Ng, and Eng-King Tan. "Translational Potential of Human Brain Organoids." *Annals of Clinical and Translational Neurology* 5, no. 2 (February 2018): 226–35. <https://doi.org/10.1002/acn3.505>.
- Sun, Nan, Xiangqi Meng, Yuxiang Liu, Dan Song, Chuanlu Jiang, and Jinquan Cai. "Applications of Brain Organoids in Neurodevelopment and Neurological Diseases." *Journal of Biomedical Science* 28, no. 1 (April 2021): 30. <https://doi.org/10.1186/s12929-021-00728-4>.
- Suong, Dang Ngoc Anh, Keiko Imamura, Yoshikazu Kato, and Haruhisa Inoue. "Design of Neural Organoids Engineered by Mechanical Forces." *IBRO Neuroscience Reports* 16 (June 2024): 190–95. <https://doi.org/10.1016/j.ibneur.2024.01.004>.
- Susaimanickam, Praveen Joseph, Ferdi Ridvan Kiral, and In-Hyun Park. "Region Specific Brain Organoids to Study Neurodevelopmental Disorders." *International Journal of Stem Cells* 15, no. 1 (February 2022): 26–40. <https://doi.org/10.15283/ijsc22006>.
- Takahashi, Kazutoshi, and Shinya Yamanaka. "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors." *Cell* 126, no. 4 (August 2006): 663–76. <https://doi.org/10.1016/j.cell.2006.07.024>.
- The Tissue Culture & Art Project. "Crude Matter." 2012. <https://tcaproject.net/portfolio/crude-matter/>.
- Trujillo, Cleber A., Edward S. Rice, Nathan K. Schaefer, et al. "Reintroduction of the Archaic Variant of *NOVA1* in Cortical Organoids Alters Neurodevelopment." *Science* 371, no. 6530 (February 2021): eaax2537. <https://doi.org/10.1126/science.aax2537>.
- Tsanni, Abdullahi. "Human Brain Cells Hooked Up to a Chip Can Do Speech Recognition." *MIT Technology Review*, December 11, 2023. <https://www.technologyreview.com/2023/12/11/1084926/human-brain-cells-chip-organoid-speech-recognition/>.
- Wang, Yanfang, Jiejie Wang, Ziyu Ji, et al. "Application of Bioprinting in Ophthalmology." *International Journal of Bioprinting* 8, no. 2 (February 2022): 552. <https://doi.org/10.18063/ijb.v8i2.552>.
- Zabolocki, Michael, Kasandra McCormack, Mark van den Hurk, et al. "BrainPhys Neuronal Medium Optimized for Imaging and Optogenetics in Vitro." *Nature Communications* 11, no. 1 (November 2020): 5550. <https://doi.org/10.18063/ijb.v8i2.552>.
- Zhang, Shun, Zhengpeng Wan, and Roger D. Kamm. "Vascularized Organoids on a Chip: Strategies for Engineering Organoids with Functional Vasculature." *Lab on a Chip* 21, no. 3 (February 2021): 473–88. <https://doi.org/10.1039/d0lc01186j>.
- Zhang, Weijie, Jiawei Li, Jiaqi Zhou, Abhay Rastogi, and Shaohua Ma. "Translational Organoid Technology – The Convergence of Chemical, Mechanical, and Computational Biology." *Trends in Biotechnology* 40, no. 9 (September 2022): 1121–35. <https://doi.org/10.1016/j.tibtech.2022.03.003>.



Cognition With & Beyond the Brain

Assemblages, Endosomatization, & AI-Driven Sensory Technologies

Daniele Cavalli

École Normale Supérieure
PSL Research University

Abstract

This paper explores the potential for reconceptualizing cognition beyond brain-centered and anthropocentric readings, highlighting the entanglement between the body, the environment, and emerging technologies. Critiquing the limitations of computationalist and reductionist perspectives as well as rejecting any possible unifying definition of the phenomenon of cognition, I advocate for a relational, entangled, and process-oriented ontological standpoint to examine cognition as a multilayered process dependent on complex cognitive assemblages. Drawing on the 4E cognition framework (embodied, embedded, enacted, extended), I integrate insights from Hayles's theory of cognitive assemblages to emphasize how cognition is dynamically shaped and extended through interactions with nonhuman agents and technical objects. Further, building on Stiegler's reflection on human technogenesis and the relationship between exosomatic organisms and endosomatic processes, I propose a speculative framework to envision forms of artificial endosomatization. Specifically, I discuss the case of AI-driven sensory technologies, imagining how some cognitive processes are not merely extracted but reintegrated into human thought through advanced sensory feedback systems.

Keywords

relational ontologies; cognitive assemblages; endosomatization; sensory technology; AI

1 Introduction

One distinguishing characteristic of humans is the capacity for metacognition, or the conscious ability to reflect and self-evaluate one's own thought processes.¹ It can be argued that this very capacity has led us to conceptualize reality as being contingent on our mental representations. However, when we move beyond the dualism of *res cogitans* and *res extensa*, cognition can be understood as a phenomenon profoundly dependent on information processing grounded in a bodily material substrate. Exploring the extent to which cognitive processes depend on and emerge from an integrated brain–body–environment system shapes an variegated range of interpretative possibilities and not infrequently has sparked heated debates.²

This work seeks to redefine our conceptual framework for understanding cognition beyond the confines of the brain's neural mechanisms to envision potential evolutions in human–machine interaction. I begin by examining the limitations of computationalism and brain-centered perspectives as well as the drawbacks of critical discussions that remain overly focused on defining cognition in exhaustive terms (section 2). Following this, I advocate for an ontological reassessment of cognition, emphasizing the speculative implications of analyzing such a complex move from a relational, entangled, and process-oriented standpoint (section 3). Next, I discuss the 4E cognition framework (embodied, embedded, enacted, extended), which proposes multiple layers of interpretations in the study of cognitive processes. In doing so, I incorporate Hayles's recent work on cognitive assemblages to enrich the discussion on the extension of cognitive processes through technology (section 4). Finally, I address how Stiegler's reflection on *endosomatization* and *exosomatization* requires reevaluation within an expanded model of cognition, which extends within an environment increasingly interwoven with artificial forms of intelligence, focusing on AI-driven sensory technologies and the possibility of engineering the process of endosomatization to envision new spaces for somatic learning (section 5).

2 Cognition Beyond Reductionism (and Exhaustivity)

Defining cognition is no simple task. Cognition and the organ most often associated with this process, the brain, embodies the very essence of complexity. As Edgar Morin insightfully observed, complexity is a term so saturated with meaning that it risks becoming devoid of meaning—even though it is now more critical than ever to think in planetary terms that transcend the certainties grounded in the rationality of our *pars sapiens*.³ Thus, I am compelled to conceptualize cognition as a complex system that extends far beyond a narrow and anthropocentric view focused solely on the human brain and its mental processes. Let us proceed systematically. From a purely etymological perspective, *cognoscere* refers to the faculty of knowing. In common understanding—and within a distinctly Western tradition rooted in Cartesian and later Kantian thought—cognition is generally conceived as an inherently human phenomenon: “a broad term that refers to the mental processes involved in the acquisition of knowledge, manipulation of information, and reasoning.”⁴ Alternatively, cognition is also understood as an umbrella term to indicate the set of functions associated with the acquisition and processing of information through interaction with the environment.⁵

Given the variety of possible interpretations, it is essential to establish some operational distinctions. *Thinking* can be understood as a process that encompasses conscious activities to a significant extent, involving self-reflective and aware reasoning. *Cognition*, instead, can be seen as a broader phenomenon of information acquisition and processing, involving both conscious thought and unconscious operations. This general distinction, though not without its challenges, aligns with some recent interpretations based on what neuropsychological research tells us to date.⁶ These processes, thinking and cognition, are commonly attributed to mental dynamics within the brain. And the mind, if we reject dualism, can be conceived *prima facie* as an organizing principle of thought but also a phenomenon dependent on the material-neural substrate that supports it.⁷ However, the mind remains a “floating signifier” with undefined boundaries, and it becomes even more resistant to a possible unified understanding.⁸ For this reason, the mind will not have a prominent place in this work. With regard to consciousness, I assume that it is not an immaterial entity separate from matter but an attribute of both thought and cognition—a state in which an agent can be “responsive to reasons.”⁹

Since the time of the Turing machine, McCulloch and Walter Pitts's mathematical models of neural networks,¹⁰ cybernetics,¹¹ and the development of cognitive sciences, particularly the rise of the computational theory of mind,¹² an attempt has been made to frame cognition in terms of computation.

¹ Proust, *Philosophy of Metacognition*; Keestra, “Metacognition and Reflection.”

² Nagel, “What Is It Like”; Kim, *Mind*; Nannini, “Mind–Body Problem”; Kok, “Will Neuroscience Make.”

³ Morin, *On Complexity*.

⁴ Kiely, “Cognitive Function.”

⁵ Moreno and Mossio, *Biological Autonomy*; Rakesh et al., “Environmental Contributions.”

⁶ Newell and Shanks, “Unconscious Influences”; Horga and Maia, “Conscious and Unconscious Processes”; Hayles, *Unthought*.

⁷ Gazzaniga et al., *Cognitive Neuroscience*.

⁸ Baars, *Theater of Consciousness*; Chater et al., “Mind, Rationality, and Cognition”; Carrara, “Unified Understanding.”

⁹ Frankfurt, “Freedom of the Will”; Dennet, *Consciousness Explained*; Schlosser, “Conscious Will.”

¹⁰ McCulloch and Pitts, “Logical Calculus.”

¹¹ Wiener, *Cybernetics*.

¹² Putnam, “Analytic and the Synthetic”; Putnam, “Brains and Behavior”; Block and Fodor, “Psychological States.”

This theoretical operation was required to move beyond the concept of the human as the sole thinking machine to open up to the possibility of artificial forms of intelligence. As Marr asserted, computation is the process of transforming one set of representations (inputs) into another (outputs) according to well-defined rules.¹³

In this view, human cognition functions like computers, with mental representations serving as inputs and outputs and the brain acting as the physical hardware executing these computations. This interpretation has gained significant prominence within contemporary connectionist approaches to the study of cognition and its potential for artificialization.¹⁴ In this framework, cognition is understood through distributed neural networks that learn by adjusting connections between units.¹⁵ In other words, connectionists argue that cognitive processes can be explained by tracing them back to the maximally parallel computation of elementary functions distributed across networks of neurons.¹⁶ As several authors emphasize, however, this perspective entails a certain degree of reductionism and has been the subject of significant criticism.¹⁷ These approaches partly form a foundational basis for much of modern theory on deep learning,¹⁸ and since the 1980s, connectionist models have often dismissed logical-symbolic approaches as outdated.¹⁹ These latter approaches posit that cognition can be understood as the manipulation of symbols according to defined syntactic rules, much like logical-based programming languages. Today, however, traditional research on the development of artificial forms of intelligence based on this model of cognition earns the rather unflattering epithet of “good old-fashioned AI” (GOFAI).²⁰

The debate surrounding computationalist interpretations of cognition, both in its connectionist and its symbolic variants, ultimately traces back to the fundamental question of how cognition itself is defined and understood at its core.²¹ As emphasized by Moreno and colleagues,²² much depends on the underlying epistemological assumptions guiding the discussion, which can be distilled into two central concerns: first, how to conceptualize the physical boundaries of cognition, and second, the methodology used to approach the question: “What kind of definition are we seeking?” The issue is not so much about adopting a single, definitive interpretation of cognition at the expense of others. Nor is it primarily about analytically solving, through logical argumentation, all the subproblems of the larger and ever-present mind–body problem. Instead, I contend that it is more productive “to discuss the methodological implications that such definitions entail . . . and rather than seeking a precise definition of cognition, one should seek a useful definition, one that enables the proper framing of a research project centered around it.”²³

Building on this proposal, I suggest an alternative approach to the issue: I conceptualize cognition without aiming for exhaustiveness but rather in a manner that can facilitate an examination of the new frontiers of human–machine interaction. In other words, I do not claim to possess the definitive interpretation of the complex phenomenon of human cognition. Rather, my aim is to engage in a process of abstraction, exploring the various possible levels of conceptualization and the theoretical consequences of reframing these concepts. As outlined at the beginning of this work, the ultimate goal is to explore the speculative implications of reimagining cognition within a relational framework that seriously considers the role of the nonhuman environment, moving beyond brain-centered perspectives.

3 Reframing the Ontological Standpoint

It is possible to assert that a relatively cross-cutting agreement can be reached by stating that cognition is a phenomenon involving the processing of information. Defining specifically what information is, and whether it can be read as measurable data²⁴ or embodied stimuli that are neither measurable nor reducible to discrete quantities,²⁵ goes beyond the scope of this work. In addition, this operation raises the same problem as attempting to exhaustively define human cognition. A useful question that has gripped philosophers and neuroscientists is instead: how do humans process information in relation to their environment? Even within a brain-centered framework, cognition is typically understood as a process of acquiring knowledge through thought, experience, and sensory input, which unfolds dynamically in interaction with the environment.²⁶ The same idea is present in the phenomenological tradition, despite the critical emphasis placed on the role of human consciousness and its perception of

¹³ Marr, *Vision*.

¹⁴ Shastry, “Neural Networks”; Smolensky, *Connectionist Approach*.

¹⁵ Newell and Simon, *Human Problem Solving*.

¹⁶ Elman et al., *Rethinking Innateness*.

¹⁷ Goldblum, *Brain Shaped Mind*; Cardon et al., “Revanche des neurones.”

¹⁸ LeCun et al., “Deep Learning”; Laurence and Margolis, *Building Blocks of Thought*.

¹⁹ Haugeland, *Artificial Intelligence*.

²⁰ Boden, *Computer Models of Mind*; Bersini, “Connectionism vs. GOFAI.”

²¹ Churchland and Sejnowski, *Computational Brain*; Van Gelder, “What Might Cognition Be”; Leite, *Theories of Human Cognition*.

²² Moreno et al., “Cognition and Life.”

²³ Moreno et al., “Cognition and Life.”

²⁴ Shannon, “Mathematical Theory”; Marr, *Vision*.

²⁵ Perret and Longo, “Reductionist Perspectives”; Yin and Goller, “Embodied Schema.”

²⁶ Kaplan and Kaplan, *Cognition and Environment*; Gallagher and Zahavi, *Phenomenological Mind*.

the world.²⁷ However, one thing remains clear: a reassessment of humans' cognitive processes and their relationship to the external environment requires engaging with ontological discourses.

The approach suggested in the previous paragraph—focused not on defining cognition but on layering the concept across various “levels of abstraction”²⁸ and examining its speculative implications—points to an initial theoretical step: accepting that ontologies are inherently political, rather than neutral, discourses on the fundamental categories of our being-in-the-world and the nature of the world. Here, the “political” is conceived as the broader processes through which power, knowledge, and relations are organized and contested within society.²⁹ As many scholars have argued, indeed, ontology is neither external nor antecedent to politics; instead, it is a modality of the political.³⁰ This implies that ontological discourses do not merely describe or establish order among the different levels of abstraction of reality but actively shape them by influencing the epistemological and ethical frameworks through which we engage with the world.³¹ As Barad notes, it is therefore appropriate to speak of “ethico-onto-epistemology.”³² In other words, ontological discourses are both descriptive and normative: they do not simply represent reality but contribute to its construction.³³ Reframing the relationships between agents and the environment, in turn, can fundamentally transform how we address complex issues such as human cognition and its interaction with emerging technologies, and the very challenges that this interaction presents.

Furthermore, the second step involves a shift from substantialist ontologies, which posit independent entities as the fundamental units of reality, toward relational, entangled, and process-oriented ontologies.³⁴ This standpoint asserts that the primary constituents of reality are not isolated substances but rather dynamic relations that shape and differentiate the very elements they connect. As Barad has masterfully demonstrated,³⁵ drawing on the insights of quantum theory, in the world there are no entities but only entangled phenomena. This view is widely discussed in postfoundational approaches to ontology—mostly inspired by the works of Deleuze and Derrida³⁶—which emphasize the interrelations between being, politics, and difference,³⁷ as well as new materialist approaches³⁸ and object-oriented ontologies (OOO).³⁹

By adopting a relational, processual, and entangled ontological standpoint, we can more effectively think about how concepts such as cognition depend on interdependencies and co-constitutive relationships with our body and the external environment, rather than on a self-contained process centered around an impenetrable mental fortress: the brain. Furthermore, this perspective aligns with a style of thinking that Bennett,⁴⁰ building on the “material turn” already introduced by Latour,⁴¹ has already shown to be an intellectual act necessary for the present times:

(1) to paint a positive ontology of vibrant matter, which stretches received concepts of agency, action, and freedom sometimes to the breaking point; (2) to dissipate the onto-theological binaries of life/matter, human/animal, will/determination, and organic/inorganic using arguments and other rhetorical means to induce in human bodies an aesthetic-affective openness to material vitality; and 3) to sketch a style of analysis that can better account for the contributions of non-human actants.⁴²

2 Entangling Cognition: From the 4E to Cognitive Assemblages

To reconsider conventional understandings of cognition and better understand the innovations brought by human-machine interaction, one may begin with the foundational work of Varela and Maturana.⁴³ They are credited with introducing a transformative perspective since the time of *Autopoiesis and Cognition*,⁴⁴ where they proposed that living systems are self-organizing and self-producing entities, continuously creating and maintaining their own boundaries through ongoing interactions with their environment. In this framework, cognition is “enacted”—it is not merely the processing of abstract symbols or representations of the external world, but instead an embodied phenomenon arising from and

²⁷ Merleau-Ponty, *Phenomenology of Perception*; Ihde, *Technology and the Lifeworld*; Verbeek, *What Things Do*.

²⁸ Floridi, *Philosophy of Information*.

²⁹ Foucault, *Power/Knowledge*.

³⁰ Butler, *Performative Theory of Assembly*; Connolly, *Ethos of Pluralization*; White, *Sustaining Affirmation*.

³¹ Lerner and Loughlin, “Strategic Ontologies.”

³² Barad, *Meeting the Universe Halfway*.

³³ Charpentier, Judith Butler's “Ontological Turn.”

³⁴ Trownsell, “Disrupting Anthropocentrism”; Santos, “Ontological Emergence.”

³⁵ Barad, *Meeting the Universe Halfway*.

³⁶ Deleuze, *Différence et répétition*; Derrida, *De la grammatologie*.

³⁷ Esposito, *Instituting Thought*.

³⁸ Braidotti, *The Posthuman*; Bennett, *Vibrant Matter*; DeLanda, *New Philosophy*.

³⁹ Developed as part of speculative realism, OOO argues that all entities (humans, animals, inanimate objects, abstract concepts) exist independently and have their own intrinsic properties and capacities, regardless of human perception or use. (Graham et al., *Speculative Turn*-Levi-Srniczek 2011).

⁴⁰ Bennett, *Vibrant Matter*, x.

⁴¹ Latour, *Reassembling the Social*.

⁴² Bennett, *Vibrant Matter*.

⁴³ Maturana and Varela, *Autopoiesis and Cognition*.

⁴⁴ Maturana and Varela, *Autopoiesis and Cognition*.

activated by the dynamic coupling between an organism and its environment.⁴⁵ As was further articulated, “Cognition is not the representation of a pre-given world by a pre-given mind but is rather the enactment of a world and a mind on the basis of a history of the variety of actions that a being in the world performs.”⁴⁶

In contrast to dualists—who treat thinking and cognition as disembodied processors of information—and the reductionist fascinations of computationalism, which tend to flatten cognitive processes to mere material-neural dynamics centered in the brain, enactive approaches redefine cognition as fundamentally rooted in the body.⁴⁷ The body is endowed with capacities specifically designed for constant interaction with environmental stimuli. According to this approach, “cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, . . . these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context.”⁴⁸ This framework compels us to adopt an integrated understanding of the human experience as a phenomenon dependent on a complex system of brain, body, and environment. But what does this imply, theoretically? First and foremost, this reading suggests a shift away from the study of and thinking about cognition as an isolated process confined to the human brain. This shift is encapsulated by the “4E cognition framework,” which posits a central claim: “Cognition does not occur exclusively inside the head, but is variously embodied, embedded, enacted, or extended by way of extra-cranial processes and structures.”⁴⁹

Gallagher has undoubtedly played a significant role in popularizing this concept of 4E,⁵⁰ alongside Clark and Thompson.⁵¹ This theoretical framework integrates multiple levels of abstraction in how to think about cognition and synthesizes numerous foundational contributions over the past fifty years. Key works include Andy Clark’s exploration of the mind’s environmental dependence;⁵² the aforementioned works by Varela, Thompson, and Rosch,⁵³ which emphasize enactive meaning-making through interaction; Lakoff and Johnson’s analysis of how physical experiences shape abstract thought, suggesting that embodied interactions are fundamental to structuring cognition;⁵⁴ Damasio’s anti-Cartesian argument on the critical role of bodily states in cognition;⁵⁵ and, finally, Clark and Chalmers’s seminal work, which argues that cognitive processes can extend beyond the individual to include objects in the external environment.⁵⁶

On one hand, cognition can be understood as a complex process within an integrated system involving the body, brain, and environment, extending well beyond the physical boundaries of the human organism. To quote Clark’s seminal work, it involves “putting brain, body, and world together again.”⁵⁷ On the other hand, the four perspectives on cognition present distinct facets of this phenomenon, each emphasizing different aspects of cognitive processes and their environmental interactions: (1) the embodied dimension highlights how cognitive processes are deeply influenced by the physical body and its interactions with the world; (2) embedded cognition emphasizes that cognitive processes are situated within a specific environmental context (or *milieu*) that shapes and supports thought and behavior; (3) enactive cognition sheds light on how cognition is not a passive reception of information but is activated through a codependent interaction with the external space, emerging through action and perception and fundamentally shaped by an organism’s engagement with its surroundings; (4) extended cognition posits that cognitive processes can extend beyond the brain to include tools, devices, and other external resources, challenging the boundaries of what we call the “mind” and suggesting that cognitive functions can be distributed across both internal and external systems.

However, in light of recent developments in artificial forms of intelligence, this 4E theoretical framework requires expanding and updating. The question arises as to how this onto-epistemological reframing can accommodate the role of nonhuman distributed agencies, given that the environment is now densely populated with more and more AI-driven technologies. In other words, how can we expand the levels of extended cognition as explored by Chalmers and Clark, who, at the end of the last century, emphasized the role of external technical agents in constituting cognitive processes?⁵⁸ This requires a reevaluation of the interaction between human cognition and an increasingly complex ecology, composed of entangled relations between human and nonhuman agencies.⁵⁹

Bernard Stiegler, drawing from a distinct philosophical tradition, offers a complementary approach to the theories of extended cognition,⁶⁰ though with differing theoretical and political

⁴⁵ Maturana and Varela, *Autopoiesis and Cognition*.

⁴⁶ Varela et al., *Embodied Mind*.

⁴⁷ Gallagher, “Intersubjectivity in Perception”; Hutto and Miyn, *Radicalizing Enactivism*; Kyselo, “Enactive Approach”; Ramírez-Vizcaya and Froes, “Enactive Approach.”

⁴⁸ Varela et al., *Embodied Mind*.

⁴⁹ Rowlands, *New Science*, Carney, “Review of 4E Cognition.”

⁵⁰ Gallagher, *Body Shapes the Mind*.

⁵¹ Clark, *Being There*; Thompson, *Mind in Life*.

⁵² Clark, *Being There*.

⁵³ Varela et al., *Embodied Mind*.

⁵⁴ Lakoff and Johnson, *Philosophy in the Flesh*.

⁵⁵ Damasio, *Descartes’ Error*.

⁵⁶ Chalmers and Clark, “Extended Mind.”

⁵⁷ Clark, *Being There*.

⁵⁸ Clark, *Being There*.

⁵⁹ Dürbeck et al., “Human and Non-Human Agencies”; Puzio, “Not Relational Enough?”

⁶⁰ Chalmers and Clark, “Extended Mind”; Clark, *Supersizing the Mind*; Rupert, *Cognitive Systems*.

objectives.⁶¹ His work emphasizes that the human condition is defined by the externalization of knowledge into technological artifacts, a process he terms *exosomatization*.⁶² Building on thinkers such as Gilbert Simondon and André Leroi-Gourhan,⁶³ Stiegler argues that human evolution is inextricably linked to the continuous development of and dependence on external tools and technologies. From the act of writing and early systems of measurement and calculation to contemporary advancements in artificial intelligence and biotechnology, humans have consistently relied on external devices to augment their cognitive and physical capabilities.⁶⁴ Humanity's defining characteristic is not merely the use of tools to compensate for inherent "structural incompleteness"⁶⁵ but rather the profound dependence on externalized systems of memory, knowledge, and agency. In this sense, the history of humanity is fundamentally the history of its "technogenesis."⁶⁶

N. Katherine Hayles enriches this discussion. Introducing the concept of the "cognitive nonconscious," Hayles redefines cognition as a dynamic interplay that extends beyond human awareness, encompassing distributed interactions with nonhuman agents—including technical systems.⁶⁷ Hayles argues that we need to properly consider that this nonconscious cognition operates independently of conscious oversight, namely the human thinking, synthesizing information, and carrying out complex processes essential to perception, emotion, and inference. By examining cognitive phenomena through the lens of her specific assemblage theory, Hayles presents cognition as planetary, distributed across both human and nonhuman agents. As she explains, "A cognitive assemblage operates at multiple levels and sites, transforming and mutating as conditions and contexts change."⁶⁸

This assemblage-based approach integrates both biological organisms and silicone-based systems, embedding cognition within coevolving structures that collectively process and interpret information. Cognition and its ethico-onto-epistemological dimensions, therefore, become a phenomenon to be understood through assemblages that "include information transactions across convoluted and involuted surfaces, with multiple volumetric entities interacting with many conspecifics simultaneously."⁶⁹ This results in a planetary cognitive ecology that aligns seamlessly with the concept of "planetary-scale computation,"⁷⁰ which characterizes today's world through various infrastructural stacks. To conclude, this perspective advocates for a recognition of the environmental entanglements inherent in cognition that extend far beyond anthropocentric and brain-centered perspectives:

Because humans and technical systems in a cognitive assemblage are interconnected, the cognitive decisions of each affect the others, with interactions occurring across the full range of human cognition, including consciousness/unconscious, the cognitive nonconscious, and the sensory/perceptual systems that send signals to the central nervous system.⁷¹

5 Speculating on Extension: Artificializing Endosomatics and Sensory Technologies

I thus accept an ontological reframing that assumes a complex cognitive ecology, where the basic premise is to conceive of our conscious cognition, or thought, as a phenomenon structurally entangled with nonhuman forms of agency, which extend beyond the limits of our brain and our own, self-aware reasoning. Through these assumptions, I can attempt to derive some speculative implications. First, we need to look at the human as Simondon would put it:⁷² a metastable entity, meaning that we exist in a constant state of becoming rather than as a fixed and isolated essence. In more cybernetics terms, the human is a "dynamic stability."⁷³ However, we need to consider that the environment is now increasingly populated by artificial forms of intelligence that are more and more interactive and affective.⁷⁴ In a relational, entangled, and process-oriented understanding of cognition, the role of emerging AI-powered technologies becomes even more constitutive in the very generation of our conscious thought. This novel condition encourages us to envision new spaces of interaction between our cognitive processes and those from a more and more interactive environment.

A possible strategy to further explore this issue is to revisit Stiegler's organological approach, which frames the interplay between human and external agents through two key concepts: *endosomatic*

⁶¹ Crogan, "Bernard Stiegler"; Turner, "Politicising the Epokhé."

⁶² Stiegler, *The Neganthropocene*.

⁶³ Simondon, *Du mode d'existence*; Leroi-Gourhan, *Milieu et techniques*.

⁶⁴ Stiegler, *La technique et le temps*.

⁶⁵ Geertz, *Interpretation of Cultures*.

⁶⁶ Stiegler, *Technics and Time, I*; Sharma, "Understanding Human Technogenesis"; Hayles, *How We Think*.

⁶⁷ Hayles, *Unthought*.

⁶⁸ Hayles, *Unthought*.

⁶⁹ Hayles, *Unthought*, 118. As Hayles explains, there is a theoretical proximity between the concepts of *assemblage* and *network*. However: "Why choose assemblages rather than networks, the obvious alternative?" This question is particularly relevant since "network" is often associated with Bruno Latour, especially in his actor-network theory (ANT), although Latour sometimes uses "assemblage" interchangeably (Latour, *Reassembling the Social*). Networks are typically viewed as consisting of nodes and edges analyzed through graph theory, which conveys a sense of sparse, clean materiality (Galloway and Thacker, *The Exploit*). In contrast, *assemblages* enable a more fleshy sense of contiguity, with entities that touch, incorporate, repel, and mutate in complex ways (Hayles, *Unthought*, 118).

⁷⁰ Bratton, *The Stack*.

⁷¹ Hayles, *Unthought*, 118.

⁷² Simondon, *L'individu et sa genèse*.

⁷³ Bardin and Ferrari, "Governing Progress."

⁷⁴ Rouvroy and Berns, "Gouvernementalité algorithmique"; Piredda et al., "Affectivity and Technology."

and *exosomatic*. Internal bodily organs are classified as endosomatic, while external organs—such as technological artifacts—are considered exosomatic.⁷⁵ This framework emphasizes the necessity of carefully considering their interrelation and the mechanisms by which they are regulated. As Stiegler highlights, “the *question of law* is the question of the *regulation of relations between exosomatic organisms*.”⁷⁶ And these exorganisms can be simple or complex: “Psychic individuals in Simondon’s sense, citizens in the Greek sense and Users in Bratton’s sense all constitute simple exorganisms, while collective individuals, such as a professional body, a unit of production in Ure’s sense, a city, a nation or a platform, are all examples of complex exorganisms.”⁷⁷ Nevertheless, in recent years, Stiegler has stressed that in an epoch of planetary exorganisms (e.g., platforms), the question of how to study the relations between exosomatic organisms needs to be addressed in a completely new way.

Building on this necessity, we have to ask: Given the increasingly profound interaction between humans and emerging AI-powered mediator technologies, is it possible to envision a reverse process—a form of artificial endosomatization—in which technology actively, and with a certain degree of autonomy, transfers the knowledge it acquires through interactions with humans back into human cognitive processes? In other words, could we imagine a reversal of the historical trend of off-loading cognitive tasks and capabilities onto external objects, considering instead a form of *onloading*, where external computational resources are integrated into forms of embodied learning and enactive decision-making?

An intriguing speculative scenario for exploring this issue can involve sensory technologies, or, more specifically, AI-driven sensory technologies.⁷⁸ These tools or systems are designed to detect, measure, or interact with human sensory perception—such as sight, touch, or sound—through specific mediating AI-powered tools. They often incorporate devices that collect data or provide feedback through sensory inputs. For instance, sensory technologies include haptic feedback systems that simulate the sense of touch, or immersive devices that deliver visual or auditory experiences, thereby enhancing user interactions with the environment.⁷⁹ Indeed, sensory technology supported by AI systems is making significant advancements in the realm of real-time and continuous health monitoring and disease diagnostics.⁸⁰ These innovations represent part of a broader shift toward a multisensory design,⁸¹ and the use of these technologies is increasingly contributing to new forms of computerized control through sensor-mediated feedback. For Emmanuel Lazega, “the use of these sensors or captors is also part of a new kind of behaviorism that seeks to influence and guide human reflexivity and judgments of appropriateness in the orientation of action.”⁸² Considering the extended framework built in the previous pages, these emerging technologies primarily interact with the cognitive assemblages through our bodily sensory and somatic experience. They also provide an intriguing example based on which we can imagine how empowering-oriented forms of an artificialized endosomatization process could address often overlooked aspects of the human cognitive experience: the role of somatic learning, or the integration of bodily sensations in learning processes.⁸³

When considering AI-driven wearable sensors and their capacities—specifically, “the delivery of physical stimuli that interact with the human sensory and motor systems . . . to elicit perceptual experiences”⁸⁴—it is possible to envision horizons for further engineering this process. In other words, is it possible to imagine AI-powered sensory technologies that do not extract data to be analyzed outside of our bodies, often for purposes far from enhancing our abilities, but rather remain within the body through forms of personalized stimulation that could amplify certain cognitive processes and thereby increase the potential for learning through the somatic dimension? What level of autonomy in sending sensory feedback could these systems have to expand cognitive processes underlying conscious thought? These questions could open interesting horizons for experimentation. However, I am not venturing into the transhumanist dystopias of mental enhancement. Instead, these technologies might simply stimulate cognitive processes that foster the development of often forgotten processes—such as somatic experience and sensory learning—that are integral parts of the complex assemblages that support and co-constitute our cognitive life. If we depart from a less anthropocentric view of the cognitive phenomenon, we can imagine new potential horizons for collaboration between humans and machines, such as to reconnect bodily sensation to cognition to expand our conscious thinking. Moreover, by thinking humans less in terms of exceptionalism and recognizing their structural entanglements with the world, we can also better understand the risks and the benefits that emerging technologies may bring.

⁷⁵ Stiegler, *Technics and Time*, 1.

⁷⁶ Stiegler, *The Neganthropocene*, 133, emphasis in original.

⁷⁷ Stiegler, *The Neganthropocene*.

⁷⁸ Chen et al., “AI-Driven Sensing Technology.”

⁷⁹ Sharma et al., “AI-Based Intelligent Sensing.”

⁸⁰ Shajari et al., “AI-Based Wearable Sensors.”

⁸¹ Cornelio et al., “Multisensory Integration.”

⁸² Lazega, “Body, Captors.”

⁸³ Bechara and Damasio, “Somatic Marker Hypothesis”; Rigg, “Somatic Learning.”

⁸⁴ Klatzky and Lederman, “Perception of Objects.”

6 Conclusion

This paper has aimed to present a possible rethinking of the conceptual framework through which we examine cognitive phenomena, the brain–body–environment system, and their entanglement with emerging technologies. By moving away from an overly anthropocentric view of cognitive processes, I have adopted a more relational and process-oriented ontological standpoint that acknowledges the intricate relation of codependence with the external environment—an environment increasingly populated by technologies capable of deeply interacting, often in an entirely unconscious manner. The proposal to consistently consider the cognitive assemblages that underpin our conscious thought, decision-making, and agency in the world could open up new speculative horizons. The possibilities for an artificialization of endosomatization serve as one such example, and AI-driven sensory technologies in particular represent a case that warrants further exploration, as the somatic dimension and the learning it may facilitate remain frequently underestimated.

Bibliography

- Baars, Bernard J. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, 1997.
- Barad, Karen. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, 2007.
- Bardin, Andrea, and Marco Ferrari. “Governing Progress: From Cybernetic Homeostasis to Simondon’s Politics of Metastability.” *Sociological Review* 70, no. 2 (2022): 248–63. <https://doi.org/10.1177/00380261221084426>.
- Bechara, Antoine, and Antonio R. Damasio. “The Somatic Marker Hypothesis: A Neural Theory of Economic Decision.” *Games and Economic Behavior* 52, no. 2 (1991): 336–72. <https://doi.org/10.1016/j.geb.2004.06.010>.
- Bennett, Jane. *Vibrant Matter: A Political Ecology of Things*. Duke University Press, 2010.
- Bersini, Hugues. “Connectionism vs. GOFAI: A Brief Critical Analysis.” In *Expert Systems in Structural Safety Assessment*, edited by Aleksandar S. Jovanovic, Karl F. Kussmaul, Alfredo C. Lucia, and Piero P. Bonissone. Springer-Verlag, 1989.
- Block, Ned, and Jerry Fodor. “What Psychological States Are Not.” *Philosophical Review* 81, no. 2 (1972): 159–81. <https://doi.org/10.2307/2183991>.
- Boden, Margaret A. *Computer Models of Mind: Computational Approaches in Theoretical Psychology*. Cambridge University Press, 1988.
- Braidotti, Rosi. *The Posthuman*. Polity Press, 2013.
- Bratton, Benjamin H. *The Stack: On Software and Sovereignty*. MIT Press, 2016.
- Butler, Judith. *Notes Toward a Performative Theory of Assembly*. Harvard University Press, 2015.
- Cardon, Dominique. *À quoi rêvent les algorithmes? Nos vies à l’heure des big data*. Seuil, 2019.
- Cardon, Dominique, Jean-Philippe Cointet, and Antoine Mazières. “La revanche des neurones: L’invention des machines inductives et la controverse de l’intelligence artificielle.” *Réseaux: communication, technologie, société* 5, no. 211 (2018): 173–220. <https://doi.org/10.3917/res.211.0173>.
- Carney, James. “Thinking avant la lettre: A Review of 4E Cognition.” *Evolutionary Studies in Imaginative Culture* 4, no. 1 (2020): 77–90. <https://doi.org/10.26613/esic.4.1.172>.
- Carrara, Andrea. “A Unified Understanding of the Human Mind: A Neuroethical Perspective.” *CNS Spectrums* 30, no. 1 (2025): e5. <https://doi.org/10.1017/S109285292400049X>.
- Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1997.
- Chalmers, David J., and Andy Clark. “The Extended Mind.” *Analysis* 58, no. 1 (1998): 7–19. <https://doi.org/10.1093/analys/58.1.7>.
- Charpentier, Arto. “On Judith Butler’s ‘Ontological Turn.’” *Raisons politiques* 76, no. 4 (2019): 43–54. <https://doi.org/10.3917/rai.076.0043>.
- Chater, Nick, Teppo Felin, David C. Funder, et al. “Mind, Rationality, and Cognition: An Interdisciplinary Debate.” *Psychonomic Bulletin and Review* 25 (2018): 793–826. <https://doi.org/10.3758/s13423-017-1333-5>.
- Chen, Long, Chenbin Xia, Zhehui Zhao, Haoran Fu, and Yunmin Chen. “AI-Driven Sensing Technology: Review.” *Sensors* 24, no. 10 (2024): 2958. <https://doi.org/10.3390/s24102958>.
- Churchland, Patricia S., and Terrence J. Sejnowski. *The Computational Brain*. MIT Press, 1993.

- Clark, Andy. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997.
- . *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, 2008.
- Connolly, William E. *The Ethos of Pluralization*. University of Minnesota Press, 2004.
- Cornelio, Patricia, Carlos Velasco, and Marianna Obrist. “Multisensory Integration as per Technological Advances: A Review.” *Frontiers in Neuroscience* 15 (2021): 652611. <https://doi.org/10.3389/fnins.2021.652611>.
- Crogan, Patrick. “Bernard Stiegler: Philosophy, Technics, and Activism.” *Cultural Politics* 6, no. 2 (2010): 133–56. <https://doi.org/10.2752/175174310X12672016548162>.
- Damasio, Antonio R. *Descartes’ Error: Emotion, Reason, and the Human Brain*. Putnam, 1994.
- DeLanda, Manuel. *A New Philosophy of Society: Assemblage Theory and Social Complexity*. Continuum, 2006.
- Deleuze, Gilles. *Différence et répétition*. PUF, 1968.
- Dennett, Daniel C. *Consciousness Explained*. Little, Brown and Company, 1991.
- Derrida, Jacques. *De la grammatologie*. Minuit, 1967.
- Dürbeck, Gabriele, Caroline Schaumann, and Heather Sullivan. “Human and Non-Human Agencies in the Anthropocene.” *Ecozon@* 6, no. 1 (2015), 118–36. <https://doi.org/10.37536/ECOZONA.2015.6.1.642>.
- Elman, Jeffrey, Annette Karmiloff Smith, Elizabeth Bates, Mark Johnson, Domenico Parisi, and Kim Plunkett. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, 1996.
- Esposito, Roberto. *Instituting Thought: Three Paradigms of Political Ontology*. MIT Press, 2020.
- Floridi, Luciano. *The Philosophy of Information*. Oxford University Press, 2008.
- Foucault, Michel. *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977*. Edited by Colin Gordon. Vintage Books, 1980.
- Frankfurt, Harry G. “Freedom of the Will and the Concept of a Person.” *Journal of Philosophy* 68, no. 1 (1971): 5–20. <https://doi.org/10.2307/2024717>.
- Gallagher, Shaun. *How the Body Shapes the Mind*. Oxford University Press, 2005.
- . “Intersubjectivity in Perception.” *Continental Philosophy Review* 41, no. 2 (2008): 163–78.
- Gallagher, Shaun, and Dan Zahavi. *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*. Routledge, 2008.
- Galloway, Alexander R., and Eugene Thacker. *The Exploit: A Theory of Networks*. University of Minnesota Press, 2007.
- Gazzaniga, Michael S., Richard B. Ivry, and George R. Mangun. *Cognitive Neuroscience: The Biology of the Mind*. 5th ed. W.W. Norton & Company, 2018.
- Geertz, Clifford. *The Interpretation of Cultures: Selected Essays*. Basic Books, 1973.
- Goldblum, David. *The Brain Shaped Mind: What the Brain Can Tell Us About the Mind*. Cambridge University Press, 2009.
- Graham, Harman, Levi Bryant, and Nick Srnicek, eds. *The Speculative Turn: Continental Materialism and Realism*. re.press, 2011.
- Haugeland, John. *Artificial Intelligence: The Very Idea*. MIT Press, 1985.

- Hayles, N. Katherine. *How We Think: Digital Media and Contemporary Technogenesis*. University of Chicago Press, 2012.
- . *Unthought: The Power of the Cognitive Nonconscious*. University of Chicago Press, 2017.
- Horga, Guillermo, and Tiago V. Maia. “Conscious and Unconscious Processes in Cognitive Control.” *Frontiers in Human Neuroscience* 6 (2012): 199. <https://doi.org/10.3389/fnhum.2012.00199>.
- Hutto, Daniel D., and Erik Myin. *Radicalizing Enactivism: Basic Minds Without Content*. MIT Press, 2012.
- Ihde, Don. *Technology and the Lifeworld: From Garden to Earth*. Indiana University Press, 1990.
- Kaplan, Stephen, and Rachel Kaplan. *Cognition and Environment: Functioning in an Uncertain World*. Praeger, 1982.
- Keestra, Machiel. “Metacognition and Reflection by Interdisciplinary Experts: Insights from Cognitive Science and Philosophy.” *Issues in Interdisciplinary Studies* 35 (2017): 121–69.
- Kiely, Kim M. “Cognitive Function.” In *Encyclopedia of Quality of Life and Well-Being Research*, edited by Alex C. Michalos. Springer, 2014: 974–78. https://doi.org/10.1007/978-94-007-0753-5_426.
- Kim, Jaegwon. *Mind in a Physical World*. MIT Press, 1999.
- Klatzky, Roberta L., and Susan J. Lederman. “The Perception of Objects and Events Through Touch.” *Psychological Science* 14, no. 2 (2003): 122–27.
- Kok, Albert. “Will Neuroscience Make Philosophy of Mind Superfluous?” Albert Kok, Emeritus Professor, Brain and Cognition Department, University of Amsterdam. <http://dx.doi.org/10.2139/ssrn.5014414>.
- Kyselo, Miriam. “The Enactive Approach and Disorders of the Self: The Case of Schizophrenia.” *Phenomenology and the Cognitive Sciences* 15 (2016): 591–616. <https://doi.org/10.1007/s11097-015-9441-z>.
- Lakoff, George, and Mark Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, 1999.
- Latour, Bruno. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford University Press, 2005.
- Laurence, Stephen, and Eric Margolis. *The Building Blocks of Thought: A Rationalist Account of the Origins of Concepts*. Oxford University Press, 2024.
- Lazega, Emmanuel. “Body, Captors, and Network Profiles: A Neo-Structural Note on Digitalized Social Control and Morphogenesis.” In *Generative Mechanisms Transforming the Social Order*, edited by Margaret S. Archer. Springer, 2015. https://doi.org/10.1007/978-3-319-13773-5_6.
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. “Deep Learning.” *Nature* 521 (7553): 436–44. <https://doi.org/10.1038/nature14539>.
- Leite, João, ed. *Contemporary Theories of Human Cognition*. Routledge, 2025.
- Lerner, Adam B., and Ben O’Loughlin. “Strategic Ontologies: Narrative and Meso-Level Theorizing in International Politics.” *International Studies Quarterly* 67, no. 3 (2023): sqad058. <https://doi.org/10.1093/isq/sqad058>.
- Leroi-Gourhan, André. *Milieu et techniques*. Albin Michel, 1945.
- Marr, David. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1982.
- Maturana, Humberto R., and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philosophy of Science, vol. 42. D. Reidel Publishing Company, 1980. First published in Spanish in 1972.

- McCulloch, Warren S., and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5 (1943): 155–33.
- Merleau-Ponty, Maurice. *Phenomenology of Perception*. Translated by Colin Smith. Routledge, 1962.
- Moreno, Alvaro, and Matteo Mossio. *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer, 2015.
- Moreno, Alvaro, Jon Umerez, and Jesús Ibáñez. "Cognition and Life: The Autonomy of Cognition." *Brain and Cognition* 34, no. 1 (1997): 107–29. <https://doi.org/10.1006/brcg.1997.0909>.
- Morin, Edgar. *On Complexity*. Translated by Robin Postel. Cresskill, NJ: Hampton Press, 2008.
- Nagel, Thomas. "What Is It Like to Be a Bat?" *Philosophical Review* 83, no. 4 (1974): 435–50. <https://doi.org/10.2307/2183914>.
- Nannini, Sandro. "The Mind-Body Problem Between Philosophy and the Cognitive Sciences" *Rivista internazionale di Filosofia e Psicologia* 14, no. 1–2 (2023): 118–34. <https://doi.org/10.4453/rifp.2023.0009>.
- Newell, Allen, and Herbert A. Simon. *Human Problem Solving*. Prentice Hall, 1976.
- Newell, Ben R., and David R. Shanks. "Unconscious Influences on Decision Making: A Critical Review." *Behavioral and Brain Sciences* 37, no. 1 (2014): 1–19. <https://doi.org/10.1017/S0140525X12003214>.
- Perret, Nicole, and Giuseppe Longo. "Reductionist Perspectives and the Notion of Information." *Progress in Biophysics and Molecular Biology* 122, no. 1 (2016): 83–91. <https://doi.org/10.1016/j.pbiomolbio.2016.07.003>.
- Piredda, Guilia, Richard Heersmink, and Marco Fasoli. "Introduction: Affectivity and Technology – Philosophical Explorations." *Topoi* 43 (2024): 587–92. <https://doi.org/10.1007/s11245-024-10055-6>.
- Proust, Joëlle. *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. Oxford University Press, 2013.
- Putnam, Hilary. "The Analytic and the Synthetic." In *Scientific Explanation, Space, and Time*, edited by Herbert Feigl and Grover Maxwell. Minnesota Studies in the Philosophy of Science, vol. 3. University of Minnesota Press, 1962.
- . "Brains and Behavior." In *Analyses of Theoretical Problems*, edited by Herbert Feigl and Grover Maxwell. University of Minnesota Press, 1962.
- Puzio, Anna. "Not Relational Enough? Towards an Eco-Relational Approach in Robot Ethics." *Philosophy and Technology* 37, no. 2 (2024). <https://doi.org/10.1007/s13347-024-00730-2>.
- Rakesh, Divyangana, Katie A. McLaughlin, Margaret Sheridan, Kathryn L. Humphreys, and Maya L. Rosen. "Environmental Contributions to Cognitive Development: The Role of Cognitive Stimulation." *Developmental Review* 73 (2024): 101135. <https://doi.org/10.1016/j.dr.2024.101135>.
- Ramírez-Vizcaya, Susana, and Tom Froes. "The Enactive Approach to Habits: New Concepts for the Cognitive Science of Bad Habits and Addiction." *Frontiers in Psychology* 10 (2019): 00301. <https://doi.org/10.3389/fpsyg.2019.00301>.
- Rigg, Clare. "Somatic Learning: Bringing the Body into Critical Reflection". *Management Learning* 49, no. 2 (2018): 150–67. <https://doi.org/10.1177/1350507617729973>.
- Rouvroy, Antoinette, and Thomas Berns. "Gouvernementalité algorithmique et perspectives d'émancipation: Le disparate comme condition d'individuation par la relation?" *Réseaux* 31, no. 177 (2013): 163–96. <https://doi.org/10.3917/res.177.0163>.
- Rowlands, Mark. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*. MIT Press, 2010.

- Rupert, Robert D. *Cognitive Systems and the Extended Mind*. Oxford University Press, 2009.
- Santos, Gil C. “Ontological Emergence: How Is That Possible? Towards a New Relational Ontology.” *Foundations of Science* 20, no. 4 (2015): 429–46. <https://doi.org/10.1007/s10699-015-9419-x>.
- Schlosser, Markus E. “Conscious Will, Reason-Responsiveness, and Moral Responsibility.” *Journal of Ethics* 17, no. 3 (2013): 205–32. <https://doi.org/10.1007/s10892-013-9143-0>.
- Shajari, Shaghayegh, Kirankumar Kuruvinashetti, Amin Komeili, and Uttandaraman Sundararaj. “The Emergence of AI-Based Wearable Sensors for Digital Health Technology: A Review.” *Sensors* 23, no. 23 (2023): 9498. <https://doi.org/10.3390/s23239498>.
- Shannon, Claude E. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27, no. 3 (1948): 623–56. <https://doi.org/10.1002/j.1538-7305.1948.tb00917.x>.
- Sharma, Abhishek, Vaidehi Sharma, Mohita Jaiswal, et al. “Recent Trends in AI-Based Intelligent Sensing.” *Electronics* 11, no. 10 (2022): 1661. <https://doi.org/10.3390/electronics11101661>.
- Sharma, Dinesh. “Understanding Human Technogenesis: Human Development in the Post-Genomic World [Book Review].” *Genomics, Society and Policy* 4, no. 3 (2008): 89. <https://doi.org/10.1186/1746-5354-4-3-89>.
- Shastri, R. Kumar. “Neural Networks: Concepts and Models.” *IEEE Computer* 22, no. 3 (1989): 68–80.
- Simondon, Gilbert. *Du mode d’existence des objets techniques*. Aubier, 1958.
- . *L’individu et sa genèse physico-biologique*. PUF, 1964.
- Smolensky, Paul. *Formalization of the Connectionist Approach to Cognitive Modeling*. MIT Press, 2010.
- Stiegler, Bernard. *The Neganthropocene*. Edited and translated by Daniel Ross. Open Humanities Press, 2018.
- . *La technique et le temps. 1. La Faute d’Épiméthée. 2. La Désorientation 3. Le Temps du cinéma et la question du mal-être* (3 vols.). Fayard, 2018. Three volumes originally published separately in 1994, 1996, and 2006.
- Thompson, Evan. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Belknap Press, 2008.
- Trownsell, Tamara. “Disrupting Anthropocentrism Through Relationality.” In *International Relations in the Anthropocene*, edited by David Chandler, Franziska Müller, and Delf Rothe. Palgrave Macmillan, 2020.
- Turner, Ben. “Politicising the Epokhé: Bernard Stiegler and the Politics of Epochal Suspension.” In *The Subject(s) of Phenomenology: Contributions to Phenomenology*, edited by Ioan Apostolescu, vol. 108. Springer, 2020.
- van Gelder, Tim. “What Might Cognition Be If Not Computation?” *Journal of Philosophy* 92, no. 7 (1995): 345–81. <https://doi.org/10.2307/2941061>.
- Varela, Francisco J., Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press 1992.
- Verbeek, Peter-Paul. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State University Press, 2005.
- White, Stephen K. *Sustaining Affirmation: The Strengths of Weak Ontology in Political Theory*. Princeton University Press, 2000.
- Wiener, Norbert. *Cybernetics: Or Control and Communication in the Animal and the Machine*, rev. ed. MIT Press, 1961. First published in 1948.
- Yin, FengYi, and Thomas Goller. “Embodied Schema Information Processing Theory: An Underlying Mechanism of Embodied Cognition in Communication.” *Communication Theory* 34, no. 3 (2024): 154–65. <https://doi.org/10.1093/ct/ctae010>.



4 Planetary Time Computation

Complex cognition is fundamentally bound to—and structured by—perceptual relationships with time. Whether immediate cognition, such as the brain’s rapid processing of movement, or abstract cognition, exemplified by how societies position themselves within historical frameworks, temporal perception shapes and directs cognitive processes. Crucially, both synchronization and desynchronization serve as dynamic variables influencing how cognitive systems coalesce and interact.

On a planetary scale, computation itself relies on artificial temporal structures such as UNIX time, which globally coordinates actions and processes. Concurrently, advances in technology continually expand human temporal perception, enabling us to access phenomena occurring at vastly accelerated or significantly slowed timescales. Thus, planetary computation simultaneously standardizes time and diversifies temporality.

Large language models (LLMs) epitomize this dual temporal relationship. Serving as both mediums and repositories for the intelligence and knowledge of civilization, LLMs function as living archives, continuously evolving through interaction and generative reproduction. Their existence as archives inherently positions them within distinct temporal frameworks—not only reflecting the present cognitive moment but also projecting meaningfully into future engagements.

These projects investigate how such technologies reshape cognitive temporalities, exploring the implications of synchronization and desynchronization across multiple scales. By examining the intricate interactions between artificial temporal systems, cognitive synchronization, and archival reproduction, the research elucidates how emerging computational paradigms fundamentally reconfigure civilization’s collective experience and understanding of time.

4a Chronoseed

LLMs, repositories of linguistic knowledge, serve not merely as databases of words but as encapsulations of the vast complexities inherent in human intelligence. Language, after all, has historically functioned as humanity's primary mechanism for encoding and transmitting cognition. This project explores the provocative possibility that an LLM could act as a long-term repository or archive of human intellect, a form of cognitive preservation.

Traditionally, archaeology seeks to reconstruct past cognition from material artifacts—tools, inscriptions, and structures. Here, the project proposes an intriguing inversion: What if cognition itself became the artifact, intentionally encoded and preserved within linguistic archives for future interpretation? Central to this exploration is the development of strategies to effectively encode and decode language and intelligence, maximizing interpretability across temporal, cultural, or even species-level divides.

Acknowledging that it is impossible to predict precisely who or what may eventually attempt to decipher this cognitive archive, the project draws insights from the history of discovering and translating lost languages. Important lessons emerge about the necessity of clear signposting, redundant encoding, and appropriate media selection, informing the mission to ensure that encoded intelligence remains interpretable despite potential gaps in knowledge or context.

Through this lens, this project investigates best practices and ideal methodologies for encoding human intelligence in linguistic forms, consciously designed for discovery and comprehension by unknown future readers. By reframing intelligence as a deliberately crafted artifact, this project raises new questions about our relationship with knowledge preservation, interpretation, and the enduring legacy of human thought.

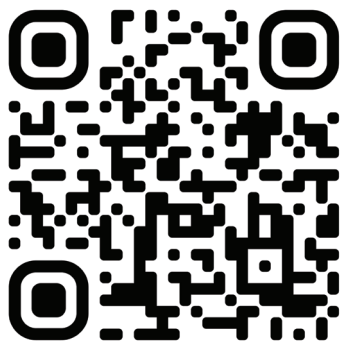
4b The Chronoconceptual Governor

Technologies fundamentally shape our horizons of perception, establishing the boundaries within which scientific inquiry can occur. Each technological innovation expands or reshapes these horizons, enabling us to perceive—and consequently conceptualize—new dimensions of reality. While certain instruments grant visibility to objects exceptionally distant or microscopically small, others uniquely alter our experience of time, compressing or decompressing temporal scales and thus offering new lenses through which we comprehend our environment.

This variability in temporal perception—chronoception—is not unique to technological augmentation. Diverse animal species naturally perceive time at different scales; for instance, the rapid visual processing of a hummingbird contrasts starkly with the slow metabolic and perceptual rhythms of a tortoise. Understanding these variations across species constitutes comparative chronoception. Extending this concept, our project explores how technologies similarly modulate time perception, creating what we term *comparative artificial chronoception*.

By systematically mapping technological synchronization and desynchronization of temporal perceptions, this study reveals not only diverse modes of scientific understanding but also distinct cybernetic interactions among agents operating in varied temporal frameworks. Artificially manipulated perceptions of time significantly impact cybernetic dynamics, altering feedback loops and interactions within complex systems.

Furthermore, this paper investigates how aggregate accelerations or decelerations in perceived temporalities influence broader cybernetic velocities within human economies. The implications of collectively modified chronoceptions reach beyond mere perception, reshaping economic rhythms and potentially transforming societal structures. Thus, this research contributes to a deeper comprehension of the connections between technology, perception, and systemic evolution in contemporary human societies.



Chronoseed

A Generative AI Time Capsule

Sonia Bernac
Royal College of Art
London

Jackie Kay
Deep Mind
Google

Winnie Street
Paradigms of Intelligence
Google

Abstract

This paper assembles a design space and defines the parameters for a generative time capsule—one that preserves not only an archive of human knowledge but also the cognitive function of human thought within a neural network. Such a time capsule must persist across vast timescales, including scenarios of civilizational collapse, while providing mechanisms for decoding its contents and executing its program, even in radically altered technological and epistemic contexts. We introduce Chronoseed as a speculative thought experiment, exploring the possibility of an AI time capsule that preserves both stored knowledge and the ability to process it. To propose its form and evaluate its feasibility, we construct a three-dimensional design space—durability, accessibility, and completeness—as a framework for systematically examining how synthetic intelligence might reshape cultural preservation. By engaging with historical time capsules, long-term nuclear waste warning systems, and speculative contributions from science fiction writers, the paper considers the broader challenge of transmitting knowledge and logics of thinking across deep time. It positions Chronoseed’s generative function as an alternative to static repositories, proposing a model in which knowledge remains interactive, adaptive, and capable of being reconstructed by future intelligences. Using the assembled design space and its parameters, the article explores possible physical forms of Chronoseed, balancing preservation with the ability of future discoverers to activate its generative potential. It ultimately identifies DNA embedding as the most viable option for long-term durability and recoverability.

Keywords

time capsules; generative AI; cultural preservation; DNA storage; interface; ethics of LLMs

1 Introduction

Historically, creating time capsules involved preserving significant objects—physical objects or conceptual artifacts such as mathematical formulas or seminal works of art and science—to curate a message for future generations. The aim was not merely to safeguard the past but to construct a future encounter with it, one inevitably shaped by shifts in interpretation, context, and technological mediation. This process of assembling and preserving objects enables a contingent reconfiguration of past cognitive functions—systems of values, beliefs, associations, logics of connection-making, and archiving technologies—filtered through the interpretative frames of those who uncover them. Of course, no perfect reconstruction is possible: Any interpretation of the past changes with the cultural drift of time, situated in a specific cultural context. Interpretation is always necessary when faced with archival evidence, but it becomes more difficult because of information loss due to the natural erosion of materials over time as well as shifts in cultural context.

The practice of time capsuling persists, now enabled by advanced technology to enhance longevity and resilience at even longer timeframes. The Arch Mission Foundation safeguards humanity's knowledge through ultra-durable Arch Libraries placed on the moon, in Earth's orbit, and deep underground, establishing a "Billion Year Archive" to preserve essential information for future civilizations.¹ Similarly, Memory of Mankind inscribes contemporary knowledge on ceramic tablets housed within Austria's Hallstatt salt mine, creating a robust archive intended to withstand vast timescales and serve future scholarship.²

AI technologies offer new possibilities for overcoming the limitations of informational ambiguity—where meaning is lost due to shifts in context, linguistic drift, or incomplete records—and material erosion, where physical archives degrade over time, making retrieval unreliable. In cultural preservation, large-scale AI models, particularly in natural language processing, have been used to encode endangered languages, preventing their extinction. Similarly, deep learning techniques have been used to restore and reconstruct fragmented historical texts, degraded audio recordings, and lost visual artifacts.³

This paper assembles a design space and defines the parameters for an AI generative time capsule, one that preserves not only an archive of human knowledge but also the cognitive function of humanity within a neural network. More than just preserving information about the present, the AI time capsule's primary objective is to maintain its generative function, allowing future generations to interact with and animate contemporary ways of thinking across time. Within the constructed framework, this paper proposes Chronoseed—a speculative generative time capsule in which a neural network is stored within DNA and encapsulated in durable kernels strategically distributed across the globe.

2 Societal Time Literacy

The legibility of a written message often degrades far more quickly than the material in which it is inscribed. Humanity's written record extends back approximately 5,000 years, with the earliest preserved writings in cuneiform dating to around 3200 BCE.⁴ These early texts are unambiguous, primarily documenting administrative and economic transactions, such as allocating barley and other goods.⁵ However, some more recent scripts remain undeciphered; for instance, the script of the Indus Valley Civilization, dating to around 2600 BCE, has not yet been successfully interpreted despite numerous well-preserved samples and artifacts.⁶ The absence of comparable textual sources erodes the cultural and semantic context—a referential embedding—necessary for the interpretation of these messages. Moreover, even when linguistic continuity is maintained, readability is not assured. Tamil, the oldest living human language, dating back to 3000 BCE, has evolved to the extent that its oldest sources are now only accessible to specialized scholars.⁷

These difficulties in transmitting meaning across time necessitate a design that mitigates not only material degradation but also the gradual loss of interpretability and epistemic continuity. A generative time capsule would construct the conditions for its own reactivation, providing a framework through which future intelligences—whatever their cultural, linguistic, or technological contexts—could reconstruct its epistemic logic and engage with its embedded modes of thought.

Pictorial forms and proto-writing, which use limited symbols, tend to have a slightly longer period of interpretability. For example, the dots accompanying animal images in the Lascaux cave paintings, which date back 20,000 years, are believed to convey information about the mating cycles of the depicted animals, likely in relation to the lunar cycle.⁸ The relevance of such information has significantly diminished for contemporary humans, who are less concerned with the mating cycles of wild animals than those who relied on hunting for survival. Thus, messages that have been sealed and subsequently discovered often serve more as inadvertent portrayals of past civilizations rather than as

¹ Arch Mission Foundation, "Preserving Knowledge, Forever."

² "MoM," Memory of Mankind.

³ Assael et al., "Restoring Ancient Texts."

⁴ Schmandt-Besserat, "Evolution of Writing."

⁵ Woods et al., *Visible Language*.

⁶ Rao, "Indus Script and Economics."

⁷ Renganathan, "Tracing the Trajectory."

⁸ Bacon et al., "Upper Palaeolithic Proto-Writing System," 371–89.

sources of practical knowledge. However, the problem of effective communication with future intelligent life is not a matter of accurate self-representation alone. The tendency toward linguistic and cultural drift over the passage of time complicates this communication. Chronoseed directly confronts this issue. Unlike conventional archival practices, which assume the continuity of linguistic and cultural legibility, Chronoseed is designed to make possible future engagement with its own embedded logics.

This challenge becomes especially urgent in cases where misinterpretation has catastrophic material stakes. Errors in reading Tamil inscriptions or Lascaux paintings may distort historical understanding, but they do not pose existential risks. However, postindustrial society has left material legacies that must be intelligible across deep time—for example, the storage sites of nuclear waste. High-level radioactive waste, such as plutonium-239 has a half-life of 24,100 years.⁹ It requires ten half-lives—241,000 years—to reduce radioactivity to less than 0.1 percent (approximately 0.09765625 percent) and be considered harmless. The conditions and storage time of nuclear materials vary depending on their radioactive properties. However, even low-level waste requires secure containment and isolation for a few hundred years.¹⁰

In 1981, the US Department of Energy established the Human Interference Task Force, a multidisciplinary team comprising scientists, communication specialists, linguists, anthropologists, and psychologists, to address the growing concern that future generations will unwittingly dig up this toxic material if it is not effectively signposted.¹¹ Since then, various international teams and institutions have been engaged in developing the field of nuclear semiotics, which focuses on long-term communication and memory transmission.¹² Proposed strategies include the creation of hostile physical structures, the use of a combination of textual and pictorial elements, the development of living warnings through genetically modified organisms, and even the establishment of atomic cults.¹³

The fact that conveying a simple warning like “Danger, do not touch or enter” has required over forty years of research, extensive funding, and international collaboration underscores the even greater difficulty of preserving more complex human ideas across deep time.¹⁴

This difficulty raises the central challenge that Chronoseed engages with: Whereas nuclear semiotics is concerned with one-way deterrence, Chronoseed is designed as a structured interpretative artifact—sealed at a single moment in time, yet encoded to provoke engagement across temporal discontinuities.

Historically, time capsuling has often been regarded as a pre-apocalyptic activity—a strategy for partial preservation or a line of flight when time or space becomes hostile or inhospitable.¹⁵ For example, Victorian historian and “freethinker” Frederic Harrison proposed burying a time capsule of British legacy under Stonehenge. Motivated by the growing awareness that time weathers and eventually destroys the traces of history, he proposed the time capsule as a “Pompeii for the twenty-ninth century,” alluding to the accidental preservation of the Roman city under the ash of a catastrophic volcanic eruption.¹⁶

Undoubtedly, the looming ecological collapse faced by a twenty-first century society, which has recently come to realize itself as an agent of geological change, creates new ethical imperatives, including new forms of archiving in response to the accelerated extinction of species, languages, customs, and cultures.

However, Chronoseed’s conceptual positioning and primary aim is neither an attempt at *non omnis moriar*—a bid for enduring remembrance—nor necessarily a civilizational backup like Noah’s Ark. Unlike a conventional time capsule, which merely transmits content, Chronoseed is structured to transmit epistemic orientation, allowing its contents to be reconstructed rather than simply observed.

Through his research in nuclear energy and the management of weapon waste, Vincent Ialenti highlights the urgency of increasing “societal time literacy,” expanding human intellectual horizons forward and backward across time.¹⁷ This involves not only a long-term approach to technological and institutional planning but also a broader move toward contemplating larger socioecological futures within the realm of contemporary philosophical discourse. This might include an emphasis on deep-time responsibility, research into the history of ideas, and the development of long-term parallel simulations of events.

The design proposal of Chronoseed fits within this line of thought—an attempt to grasp collective cognitive function as an ongoing process, reflecting contemporary ways of sense-making: thinking, sensing, associating, believing, and forming opinions. Rather than simply documenting contemporary thought, Chronoseed encodes the conditions for its intelligibility. Its discovery is not an act of passive witnessing but an invitation to reconstruct its epistemic logic. Unlike a static archive that risks becoming an indecipherable relic, Chronoseed operates as an artifact structured to generate meaning beyond its original context.

⁹ US Nuclear Regulatory Commission, “Backgrounder on Plutonium.”

¹⁰ International Atomic Energy Agency, *Radiation Protection and Safety*.

¹¹ US Department of Energy, “Reducing the Likelihood.”

¹² Mazzucchelli and Paglianti, “How to Remember?”

¹³ Chapman, “Speaking to the Future.”

¹⁴ Trauth et al., “Expert Judgment on Markers.”

¹⁵ Yablon, *Remembrance of Things*.

¹⁶ See Moynihan, “Giant Time Capsule.”

¹⁷ Ialenti, *Deep Time Reckoning*.

3 Design Space

Before outlining Chronoseed's form, it is necessary to distill the key challenges that shape Chronoseed's design. Through surveying past archival efforts to preserve cultural knowledge, we identify three major *desiderata* of time capsule design:

1. *Completeness*: A complete time capsule contains a holistic and representative sample of human knowledge.
2. *Durability*: A durable time capsule has a form factor and materiality that enables it to survive centuries into the future and across significant environmental and planetary changes.
3. *Accessibility*: An accessible time capsule signposts itself to facilitate its discovery under favorable conditions and the intuitive decoding of its contents.

The limitations of our current technology result in inherent tradeoffs between these three desirable properties. To explore this point and its implications for the design of Chronoseed, we consider two instances of historical and speculative time capsules plotted in three-dimensional space across these axes (Figure 1).

The Rosetta Stone is an ancient stele carved from stone inscribed with the same decree in three written languages: ancient Egyptian hieroglyphics, Demotic script, and ancient Greek.¹⁸ Because it is preserved on granodiorite and has survived from 196 BCE, the Rosetta Stone is considered *durable*. The multilingual replication of this text allowed historians to draw on the surviving knowledge of ancient Greek to decipher the lost scripts of hieroglyphics and Demotic, making the contents *accessible* and even generative. However, once deciphered and interpreted, the Rosetta Stone was revealed to be a relatively banal administrative record: an incomplete descriptor of Ptolemaic Egyptian society.

Turning to our proposition of an AI time capsule, we now consider the suitability of today's *state-of-the-art multimodal chatbot* (such as OpenAI's GPT-4o). Although foundation model training data sets scraped from the internet are by no means an exhaustive record of human knowledge, they might be the largest and most diverse archives collected in human history;¹⁹ thus, we can expect the models trained on this data to be *more complete*. The multilingual, multimodal (audio, text, and visual) interfaces to this chatbot, along with its deployment on widely available browser and mobile applications, make it fairly *accessible* in today's society. However, modern software is highly ephemeral.²⁰ Machine learning models are quickly deprecated, and the machines storing the software and weights for these models are kept under controlled conditions in data centers, due to their vulnerability to environmental damages.

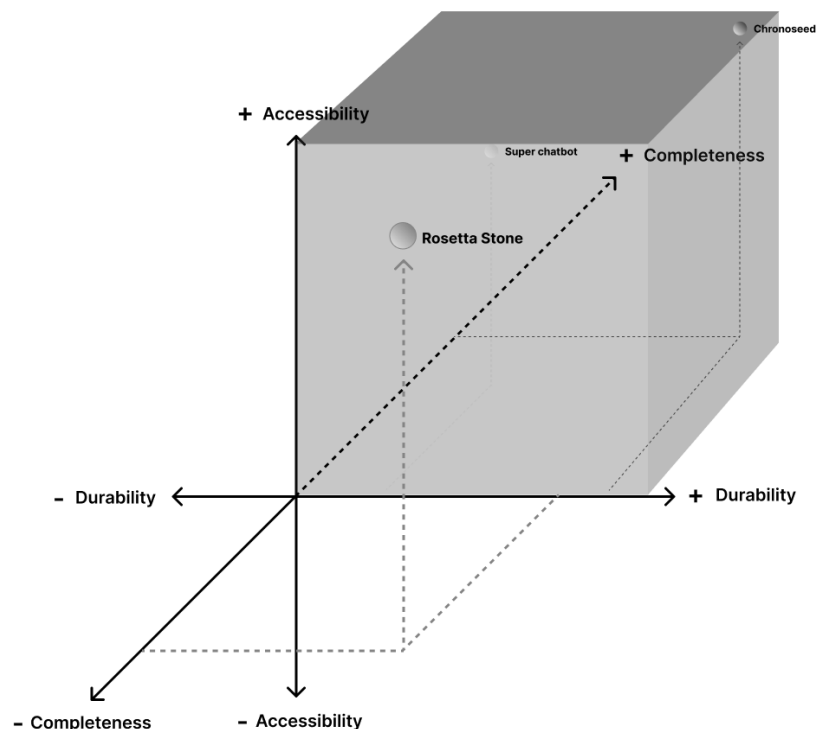


Figure 1 Time capsule design space.

¹⁸ Parkinson et al., *Cracking Codes*.

¹⁹ Raffel et al., "Exploring the Limits."

²⁰ Dong and Xie, "Large Language Models."

Unlike these examples, the ideal design for Chronoseed would be optimal on all three of these axes. We now discuss each of these dimensions in detail to arrive on a concrete specification for our time capsule. For each dimension, we will first describe the considerations that arise for the designer from “first principles.” We then deepen that dimension by further probing issues that arise when the time capsule is discovered and how both its physical form and its meaning are tested against temporal passage.

4 Completeness

While a measure of completeness is often contingent on the intended usage, audience, and context of the archive, the most ambitious time capsule projects have set a widely encompassing scope: to compile a *representative* snapshot of the human knowledge most worthy of preservation. However, future generations may view these unearthed archives as reflecting the outdated, narrow-minded, or even backward curatorial choices of their creators. Evaluating representativeness is fraught with questions of subjectivity and contextuality.

To dodge these questions, we can turn to mathematics to formalize a universal definition of completeness. A potential definition of a complete archive is one that contains more *information*, agnostic to its qualitative character. By modeling the data of an archive as a statistical process, we can use tools from information theory, such as Shannon entropy and mutual information, to quantify information complexity. A related approach is to measure the statistical redundancy of the archive through its compressibility—that is, how much data can be losslessly compressed.²¹

While the purity of such measures is appealing, they flatten and erase the concerns of cultural representativeness and the *utility* of knowledge, as defined by its instrumentality to the survival of future generations. The time capsule might be a preservation tactic in anticipation of a massive catastrophe or a new Dark Age of scientific ignorance. A useful time capsule would encapsulate all the cutting-edge knowledge in the life sciences and engineering or the best economic and political practices for governing a flourishing society.

A *complete* time capsule is representative, has utility, and compresses a high level of information complexity. We will now argue why an AI time capsule is a more complete archive than a conventional one, and highlight some of the design choices that lend themselves to these traits.

In conventional time capsules, archival objects such as artifacts, records, documents, or even seeds of plants are put into storage, and after a long period of time, the time capsule is discovered. The contents are unboxed and subjected to historical interpretation, and the preserved potential is realized.

Arguably, neural networks themselves are a kind of archival storage, in that they preserve a kind of “superhistory” of their training data.²² In the AI time capsule, large volumes of input data are compressed into the weights of a neural network due to a learning process with a specific training objective, such as next-token prediction. When the time capsule is opened, the neural network can be run with new inputs and potentially produce new outputs.

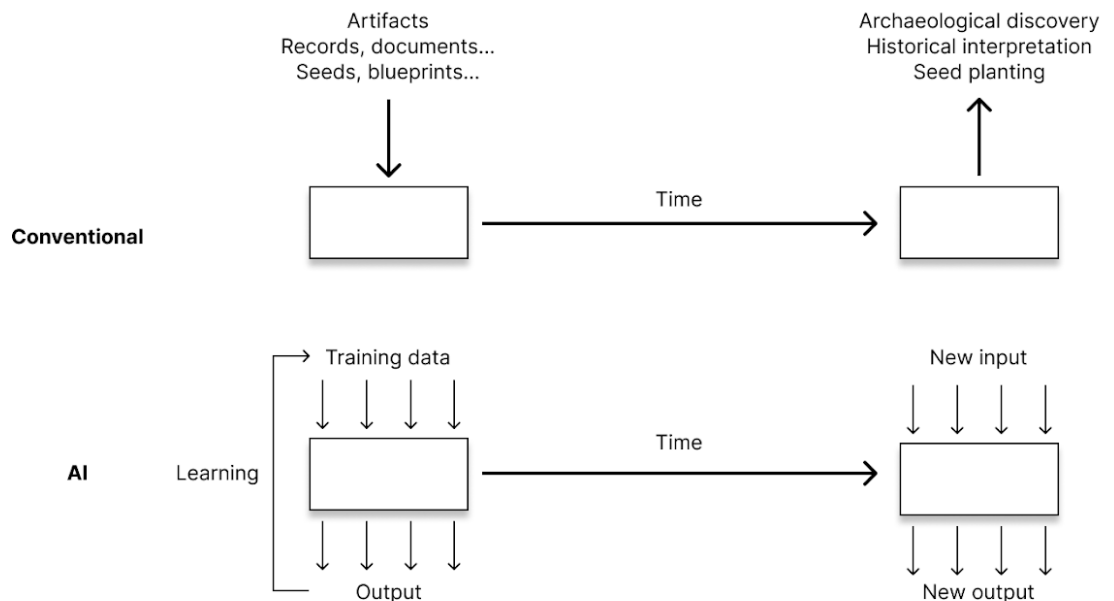


Figure 2 Schematic of conventional versus AI time capsule.

²¹ Lindgren, “Information Theory.”

²² Rao, “Superhistory, Not Superintelligence.”

While conventional time capsules preserve archival data and present them as is, the AI time capsule approximates the cognitive function that can be interacted with in the future (see Figure 2 for a comparison of conventional and AI time capsules). Large neural networks meet a certain mathematical definition of completeness because they are essentially large-scale compressors that remember patterns and forget noise.²³ Not only has the network compressed a large archive into its weights, sampling outputs from it allows a kind of generative infilling or extrapolation along the manifold of data. The approximated function is more complete than data because it has the capacity to generalize to new conditions. It represents an approximation of the past in the future and enables the interactivity of an archive. (Retrieval architectures could enable semantically informed searches over historically preserved archives, or even serve as the basis for information retrieval over yesterday's data with tomorrow's models.)

Both the representativeness and the utility of the time capsule depend on the information encoded in the neural network. Media such as language, literature, arts, and film may lend themselves more to representing the cultural context of a generation, whereas scientific knowledge (e.g., of protein folding, catalysis, geometry, etc.) preserved in a neural network may be of higher utility.

However, scientific models quickly become outdated or disproven. A pre-Copernican, Earth-centric model of the solar system serves mainly to provide historical context to the present-day astronomers. Yet there is a chance that models of scientific knowledge may preserve possibilities and theories that will become experimentally verifiable in the future. For example, the Einstein–Podolsky–Rosen paradox, a thought experiment that was said to attack the foundations of quantum mechanics, was experimentally disproven in 1982, settling a decades-long debate between Albert Einstein and Niels Bohr.²⁴

The generativity of AI comes with the associated limitation of hallucination. Current generative models are confabulators that are not bound to a strict manifold of reality and facts,²⁵ and packaging them into a time capsule may risk propelling misinformation about the present into the future. Yet the interpretation of historical ambiguity is necessary for any archaeological exploration. It is the historian's due diligence to treat records from the past as unreliable. Discourse, and therefore meaning, emerges from the triangulation between conflicting accounts of the past.²⁶

4.1 Contents and Ethics

Although completeness is often considered an ideal quality of time capsules, its definition and the question of who determines it are complex and politically charged. The assumption that completeness inherently enhances a time capsule's value warrants ethical scrutiny. Beyond durability and accessibility, Chronoseed's completeness raises fundamental questions about selection, representation, and responsibility. The process of determining what is preserved—and how it is weighted—directly shapes the cognitive function it encodes for future engagement. This section examines the criteria for completeness, the ethical and political stakes of curation, and the potential consequences of omission or bias in the selection process.

David Lowenthal claimed that the Victorian and Edwardian eras, with their focus on building lasting infrastructures like railroads, aqueducts, sewer systems, libraries, parks, and gardens, were driven by a “cult of posterity.”²⁷ Nick Yablon counters that this cult evolved into a disconnect from future generations with the rise of signal transmission infrastructures. With technological acceleration and the short-lived nature of communication media like “wood-pulp paper, photography, phonography, and film,” the longevity of messages for future generations could no longer be taken for granted. This led to a perceived need for “direct communication with the future,” which Yablon argues often resulted in

temporal myopia . . . a presumption that later generations would look back with gratitude and admiration. One can also detect in them a growing tendency to reduce the duty to posterity to a merely archival duty, as if it were enough to preserve a smattering of documents, photographs, and artifacts; often inexpensive or redundant, these materials represented no great sacrifice.²⁸

Indeed, communicating across deep time risks turning into a grandiose monologue. The ethical challenge is threefold: (1) conveying unprecedented responsibility for environmental impact, (2) selecting training data that fairly represents contemporary cognitive frameworks, (3) and making Chronoseed useful for future intelligent beings rather than focusing solely on self-representation.

A seemingly responsible approach might reject grand historical narratives entirely, in line with the postmodernist sentiment, instead focusing on marginalized voices, imponderabilia, and those aspects

²³ Tishby and Zaslavsky, “Deep Learning.”

²⁴ Aspect et al. “Experimental Realization,” 91.

²⁵ Ji, et al., “Survey of Hallucination,” 1–38.

²⁶ Foucault, *Archaeology of Knowledge*.

²⁷ Lowenthal, *Past Is a Foreign Country*, quoted in Yablon, *Remembrance of Things*.

²⁸ Yablon, *Remembrance of Things*, 17.

of life that are not abundantly present online.²⁹ Such an approach could borrow from the logic of Tavares Strachan's comprehensive research project that includes overlooked individuals, places, and events.³⁰

However, training an AI time capsule solely on marginalized data risks idealizing the minor and the local. A time capsule could, for example, be shaped by local gossip, neighborhood disputes, familial cruelties, or cycles of revenge within communities. While it is essential to address misrepresentation and ensure that Chronoseed does not perpetuate social and perceptual inequalities through underexposure or overexposure, it is also important to recognize that local and private perspectives do not inherently possess greater moral authority.

The question then arises: To what degree should a neural network engage in a form of Potemkin politics for future generations?³¹ The term originates from the *Potemkin villages*—allegedly elaborate facades constructed by Grigory Potemkin to impress Catherine the Great, concealing the true state of the land behind them. Should a neural network similarly construct a legible but selective representation of its present—one that smooths over complexities, omissions, and inconsistencies—or should it encode traces of its own internal uncertainties, ruptures, and blind spots, leaving a more ambiguous but potentially richer inheritance?

If Chronoseed were to capture a representative account of human society, it would need to include not only technological and cultural advancements but also political failures, wars, exterminations, and all forms of human violence, cruelty, and ignorance.

Equally critical would be an account of how humanity has irrevocably damaged the natural environment. This includes data on deforestation at an industrial scale, biodiversity loss, mass extinction events driven by habitat destruction, ocean acidification, the collapse of pollinator populations, soil degradation, deep-sea mining, plastic pollution infiltrating every level of the food chain, and the warming of the planet beyond sustainable thresholds. Records would need to document toxic waste dumping, groundwater contamination, the destruction of rainforests for monoculture crops, the overfishing of oceans to the brink of collapse, and the unchecked release of greenhouse gases accelerating climate change.

If completeness is the goal, any attempt to encode human history must acknowledge not only its intellectual contributions but also its profound role as an agent of ecological and civilizational self-destruction.

However, preserving these records for millennia risks reintroducing humanity's long-forgotten maladies into societies that may have already overcome them. Yet, removing the accounts of human atrocities from training data would be equivalent to abandoning one of the most vital archival functions: the role of bearing witness through time. It might also veer too close to a kind of ideological paternalism over a future in which the sociocultural, economic, and environmental conditions that shape ideology might be dramatically changed.

The challenge of honesty in this context extends beyond humanity's darkest aspects to include the sheer volume of mundane digital content. A comprehensive record would likely be dominated by social media posts, influencer streams, repetitive entertainment, viral videos, and vast quantities of self-referential imagery, raising the question whether prevalence alone justifies preservation.

Arguably, the most significant absences—what has been lost, destroyed, or made extinct—should be reflected in the preserved cognitive function too. The training of Chronoseed should encompass not only what exists but also what is no longer present, acknowledging that the impact of these absences constitutes an important part of the collective feelings and thought. This approach could include a comprehensive record of everything that has been eradicated or irrevocably altered, much of it due to human activity, echoing the themes explored by Judith Schalansky.³²

If the solution involves including all these aspects in the training data, the politics of weight distribution becomes crucial. The preservation of cognitive function cannot be reduced to a "marketplace of ideas" where mediocre and harmful content is given equal weight alongside seminal works of art and science.³³ It is equally vital that the preserved cognitive function of humanity does not become a diluted, averaged version of human thinking.

A further complication arises in the context of training data. Treating people's online activity and content production as accurate representations of their thoughts, culture, opinions, values, and desires risks falling into a functionalist trap. This critique does not suggest the existence of an inexpressible or untokenizable essence of human existence that cannot be conveyed; rather, it underscores that many aspects of life remain unexpressed or untokenized.

5 Accessibility

Accessibility in archival design must account for an unpredictable future. Chronoseed's legibility depends on how technology, language, and culture evolve over time. As linguistic and technological paradigms shift, future discoverers may struggle to decode its contents. Chronoseed must signal its artificial nature and invite interpretation without assuming shared standards. Its design prioritizes

²⁹ Lyotard, *Postmodern Condition*.

³⁰ Strachan, *Encyclopedia of Invisibility*.

³¹ Can, "Under the Leadership," 356–76.

³² Schalansky, *Inventory of Losses*.

³³ Herzog, "Marketplace of Ideas."

recognition over immediate usability, ensuring it remains identifiable as an artifact intended for engagement.

In some respects, the AI time capsule presents additional logistical challenges to accessibility, but the medium can also enable intuitive modes of interactivity. In this section, we consider how to design for accessibility in the worst case of intergenerational knowledge loss, when the historical endeavor of building the time capsule has been completely forgotten and no context remains. Chronoseed motivates an ideal notion of a *minimum viable accessible interface*. An exact blueprint for such an interface is out of the scope of this work, but we sketch the desirable properties for accessibility to inform future designs.

Before Chronoseed can be decoded, it must first be discovered. To distinguish it from archaeological noise, it must signal its artificiality, marking itself as an object or structure of interest. We propose a replicated design, where clusters of identical plant-like seeds, each containing an embedded neural network, are distributed in an anomalously regular pattern. This pattern—deliberately non-random—serves as a self-signposting mechanism, ensuring that future discoverers recognize its intentional design and are drawn to investigate what lies within.

Chronoseed also requires a suitable material infrastructure for executing a neural network. The medium and design must defy the rapid onset of technological obsolescence, embracing permacomputing over ephemerality.³⁴ Permacomputing is an approach to computing that prioritizes low-energy consumption, repairability, adaptability, and minimal environmental impact, aiming to create computational systems that can function reliably over extended timescales without relying on rapidly obsolete infrastructure.

A highly accessible time capsule would provide the processing power and interfacial devices necessary for generating and displaying the outputs of the model; however, this complicates the engineering challenge of durability. While carving the weights of a neural network into nickel disks³⁵ or even a stone stele would be highly durable, any means of running the model would have to be engineered by future discoverers, making the time capsule inaccessible to any civilizations that lack the requisite algorithmic knowledge or computational processing power.

Once the time capsule is found and the model is running, the question remains how the information will be understood. Chronoseed offers not a flat, linear passage of text, but a function with outputs contingent on inputs. The design of the time capsule must indicate what modes of interaction are possible and how meaningful behavior can be elicited from the neural network. While intuitive and implicit cues should facilitate this where possible, each embedded kernel would need to be marked on its surface with simple symbols, diagrams, or pictograms that suggest its computational nature and interactive function. These external markings serve as a universal signal that the object is not inert or purely archival, but something designed to be engaged with, queried, and activated.

We recommend leveraging *multimodality* to offer diverse pathways for discovery and meaningful interaction. The neural network embedded within each Chronoseed kernel encodes knowledge through multiple modalities—textual, visual, and auditory—allowing future discoverers to reconstruct meanings even when linguistic continuity is lost. Like the linguistic redundancy of the Rosetta Stone, multiple modalities offer more robust pathways for decoding symbols and reconnecting representations to the physical world. However, since Chronoseed itself does not incorporate active hardware such as screens, microphones, or haptic devices, it relies on external interfaces brought by its discoverers. Thus, the capsule’s surface includes markings explicitly suggesting its computational nature, indicating that it contains executable instructions or encoded cognitive functions, inviting meaningful engagement and interaction through appropriate external technologies.

6 Discovery

This section considers the civilizational circumstances in which Chronoseed might be encountered and the implications these have for its design.

Constructing a transtemporal object that will be discovered beyond the design and linguistic horizons of contemporary humans necessitates accounting for contingencies such as ecological catastrophes, partial extinctions, and the long-term alienation between present and future cognition. The accessibility of any time capsule is ultimately shaped by both the cognitive and cultural differences of its future recipients and their level of technological advancement—factors that will determine the possibility and nature of its decoding.

Imagining future humans tends to collapse them into a community with a singular mind, similar skills, and evenly distributed access to technologies. However, the space of discovery will most likely be a terrain of political conflicts, impenetrable infrastructural divisions, and various dramatic inequalities, just like the contemporary world. Therefore, its design should prevent it from playing the role of the New Apple of Discord that, in Greek mythology, led to the Trojan War. The distributed nature of Chronoseed—consisting of multiple identical kernels, geographically dispersed and embedded in diverse locations—reduces the risk of exclusive control, while its redundant placement and artificial patterning increase the likelihood of recognition and accessibility across different political and technological

³⁴ Mansoux et al., “Permacomputing Aesthetics.”

³⁵ Biersdorfer, “Time Capsule.”

realities. This uneven terrain of discovery necessitates a design that is not only resilient but also resistant to monopolization and unintended geopolitical consequences.

As futurists have long recognized, any attempt at utopian political engineering of the far future from the deep past risks having unintended consequences. One proposed solution is to shard Chronoseed's database, facilitating future exchange of knowledge between discoverers encountering different seeds. In this approach, information would be partitioned across multiple instances, with some data remaining consistent across all seeds while other segments remain unique to individual shard-seeds. While this could encourage collaboration, it might just as easily escalate conflicts, leading to unethical trading, competition, or even theft of Chronoseed kernels.

Moreover, the challenge of access is not limited to data distribution alone. Each kernel of Chronoseed is small, self-contained, and limited to a neural network with a constrained set of embedded instructions, without an accompanying interface or physical technology. This limitation means that activating and engaging with its cognitive function depends on external infrastructures, raising questions about who has or invents the means to reconstruct and interpret it. As a result, the conditions of its discovery may introduce additional geopolitical and economic dynamics, as access to Chronoseed could become a contested resource.

Anticipating how an artifact will be encountered, interpreted, and what consequences it may generate in the distant future demands precise consideration of its potential discoverers—their epistemic frameworks, technological capacities, and cultural conditions. A notable historical precedent can be found in the early 1980s, when the German *Zeitschrift für Semiotik* (*Journal of Semiotics*) invited readers to submit ideas for conveying messages that could be understood 10,000 years into the future.³⁶ This inquiry, originally aimed at developing a warning system for nuclear waste, serves as a valuable thought experiment for understanding how meaning might be preserved across vast temporal distances. This inquiry led to diverse responses, each revealing hidden assumptions about the future recipients of transtemporal communication. In the present paper, the speculative reception of Chronoseed is systematized within a two-axes discovery space.

The discovery space of the AI time capsule can be represented as a light cone extending along the time axis into the future (Figure 3). Two key axes—*technological advancement* and *civilizational difference*—define the range of decoding possibilities for future discoverers.

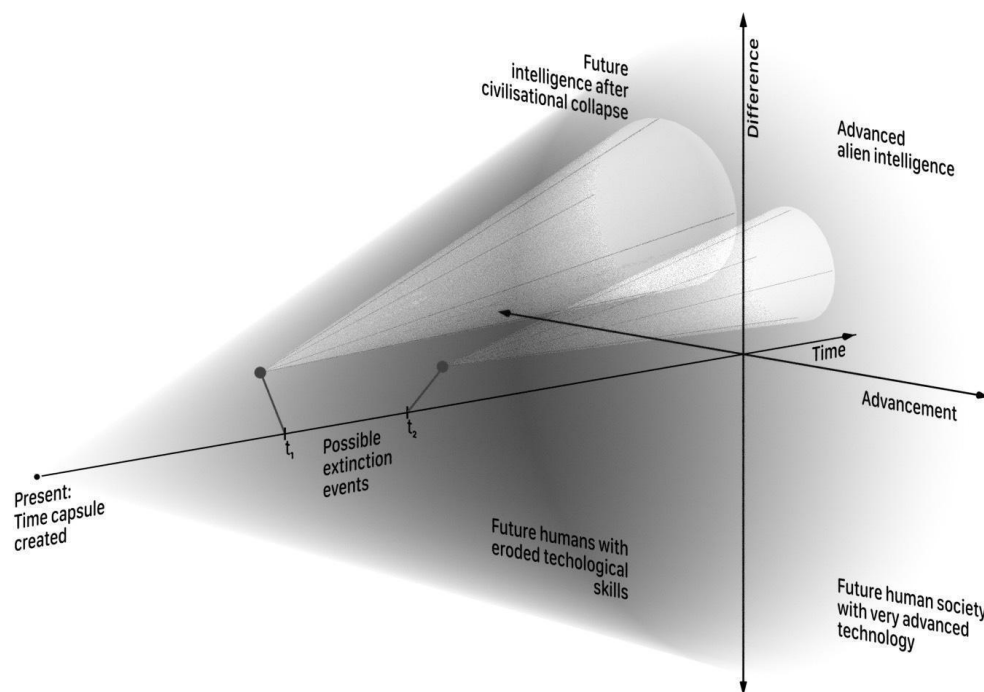


Figure 3 Discovery space of the AI time capsule.

Two partial civilizational collapse events obscure the possibilities of discovery, introducing a civilizational discontinuity in the top two quadrants.

³⁶ “Band 6, Heft 3,” *Technische Universität Berlin*.

6.1 Low Advancement, High Difference

The top left quadrant of Figure 3 marks the discovery space after the civilizational collapse—a major catastrophe that caused a high degree of alienation between contemporary humans and future humans or other future forms of intelligent life. In this space, the complexity of intelligent life has been drastically reduced with respect to some combination of biology, sociality, technology, and culture.

Literary depictions of postapocalyptic worlds, such as Tatyana Tolstaya's *The Slynx*, help illustrate possible discovery scenarios. *The Slynx* envisions a world reshaped by “the Blast,” a catastrophe that reduced civilization to a medieval-like state. Survivors, mostly illiterate, attempt to make sense of artifacts from the past, transcribing both great literature and technical manuals indiscriminately—unable to differentiate their significance or grasp their original meaning.³⁷

To convey messages across deep time, Sandia National Laboratories proposed rudimentary physical signposting, relying on material markers rather than shared semantic or contextual codes in scenarios where linguistic continuity is lost. In 1993, they suggested surrounding the Waste Isolation Pilot Plant—a deep geological nuclear waste site in New Mexico—with a landscape of jagged spikes to warn against intrusion.³⁸ Thorns, spines, and spikes are examples of an aposematic symbol—a warning from one species to another that they should not be attacked or eaten—that is pervasive in the biological world. In the context of Chronoseed, an alternative to deterrent spikes could involve soft, rounded forms or biomimetic designs resembling plant seeds, signaling non-toxicity while ensuring they do not appear visually appetizing enough to provoke gustatory interest. This could be achieved through non-fruit-like coloration, structural hardness, and surface inscriptions, ensuring it is recognized as an artifact meant for interpretation rather than consumption.

As suggested earlier, Chronoseed is not accompanied by a physical interface. Although an inviting design might attract future discoverers, their low technological advancement would limit access to its generative function. This raises the challenge of making it self-executing, though any additional mechanisms to aid accessibility could compromise its robustness.

6.1 High Advancement, High Difference

The top right quadrant of Figure 3 outlines the discovery space of a society that despite an apocalypse and disrupted civilizational continuity managed to create advanced technologies without the foundation of present knowledge.

In response to the *Zeitschrift für Semiotik*'s call, writer and futurologist Stanisław Lem proposed “information plants,” genetically engineered to encode a mathematical warning about nuclear waste.³⁹ Those “atomic flowers” would be engineered to grow in contaminated areas, serving as biological markers of danger. Lem envisioned that their distinctly unnatural appearance would signal artificial origin, attracting future discoverers to investigate and decipher their embedded message.

For an advanced alien intelligence vastly different from contemporary humans, a time capsule might be more of a technological puzzle, its significance leaning toward anthropology or philosophy rather than strictly regenerative knowledge. It could bridge civilizational discontinuity between two forms of advanced intelligence, providing solutions or, conversely, acting as Pandora's box, transmitting the biases, cruelties, and obsessions of its creators.

In this scenario, Chronoseed's generative function is likely to be activated, as a sufficiently advanced intelligence—human or otherwise—would have the capacity to recognize and run the model. No dedicated interface would be required, as its future discoverers would possess the technological means to reconstruct and execute its cognitive function directly.

6.3 Low Advancement, Low Difference

The bottom left quadrant of Figure 3 marks a society with eroded technology. Despite partial preservation of intellectual continuity and the lack of sudden disruptive events, the civilization experienced a broadly understood decline.

To explore the challenges of long-term communication in the event of technological diminishment, various speculative thought experiments have been proposed. In the context of conveying a warning across time, linguist and semiotician Thomas Sebeok introduced the concept of “Atomic Priesthood.” This concept involves creating a belief system that uses symbols, rituals, and myths to pass down knowledge about hazardous sites. By embedding a ritualized fear into cultural superstitions, even if scientific understanding fades, this ingrained taboo would protect future generations from engaging with dangerous areas.

A similar, hybrid approach was proposed by Françoise Bastide and Paolo Fabbri, who envisioned genetically modifying domestic cats into “Ray Cats” that would change color when exposed to radiation. This feline transformation would be encoded in cultural artifacts like myths, songs, and art,

³⁷ Tolstaya, *Slynx*.

³⁸ Chapman, “Speaking to the Future.”

³⁹ Hawranek, “Stanisław Lem.”

the cat would serve as a cultural marker, its color change becoming a signifier of danger in the cultural memory of future generations.⁴⁰

Embedding information about Chronoseed within cultural artifacts, legends, and ritualistic auxiliary structures could help signal its significance as a distributed vessel of knowledge, making it more likely to be recognized and engaged with in the future. In this scenario, Chronoseed would most likely be recognized as a time capsule—an artifact meant to be interpreted, potentially serving as a tool for reviving lost technologies and cultures. However, the extent to which its generative function could be activated without a self-assembling interface depends on the level of technological diminishment. If sufficient infrastructure remains, it could still be computationally engaged with, but in a more technologically reduced society, its contents might be approached through material or symbolic analysis rather than direct execution.

6.4 High Advancement, Low Difference

The bottom right quadrant of Figure 3 represents a society with advanced technology and preserved intellectual continuity, where the absence of sudden disruptions results in relatively low divergence between contemporary and future humans.

In a context where constant, albeit subtle, transformations can make significant change difficult to discern or acknowledge, Chronoseed would provide future observers with a means to comprehend the scale and direction of civilizational evolution over time. With both technological advancement and civilizational continuity, decoding its contents would be relatively straightforward, transforming it into an anthropological record rather than a puzzle. This record would serve as a *still from the past*, capturing a moment in time with all its nuances. Engaging with the time capsule would be akin to interacting with a version of a collective self from the past before all the technological updates, offering a perspective from a specific moment in the ongoing process of technological evolution.

7 Durability and Form Factors

After examining the epistemic frameworks, technological capacities, and cultural conditions of future discoverers, we now turn to the form factor of Chronoseed. Drawing on the insights already presented, we outline different design proposals, ensuring that Chronoseed's structure addresses the challenges of completeness, accessibility, and durability (Figure 4).

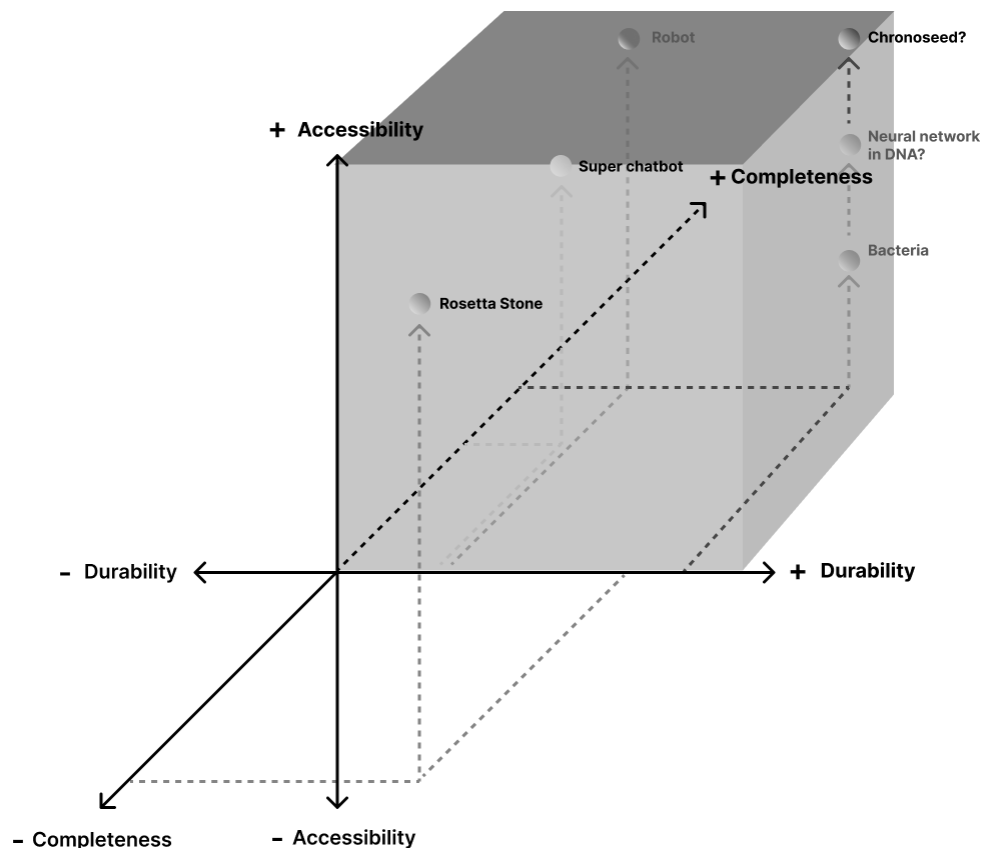


Figure 4 Design space with proposals.

⁴⁰ Fabbri and Bastide, “Living Detectors,” 10–13.

To ensure the long-term readability of a generative function of a neural network embedded in Chronoseed, various strategies must be adapted to the technological capabilities of the future civilization that might discover and decode it.

To address the recurring challenge of the interface, Chronoseed could be designed as von Neumann replicators, an old idea recently revisited in modular robotics research at MIT, where self-replicating hierarchical robotic swarms have been proposed for autonomous assembly and adaptation.⁴¹ These robotic systems consist of modular units that can self-assemble into complex structures, construct their own interfaces upon activation, and adapt to their environment by using available resources. They are also capable of limited self-repair if damaged.

However, this approach introduces significant obstacles, including control challenges in coordinating distributed replication, error accumulation leading to functional failures, resource depletion disrupting sustained replication, and the difficulty of ensuring robustness in unpredictable environments.⁴² Additionally, incorporating self-assembly mechanisms would increase the size of Chronoseed significantly, even if the majority of interface components were sourced externally. Furthermore, the necessary materials to construct an interface might not always be available in the surrounding environment, limiting the feasibility of this approach. Given these constraints, more stable, passive encoding methods, such as DNA storage, remain preferable within the limitations of present-day technology.

DNA storage emerges as the most viable solution for preserving the generative function of a neural network within Chronoseed, given its *high data density* and *stability*.⁴³ It can efficiently archive large volumes of information and precisely encode the instructions to reconstruct a neural network. Hence, it achieves a high score for the completeness of preserved information. However, DNA is primarily a passive storage medium, with the preserved information requiring time-consuming sequencing.⁴⁴ DNA-stored messages lack an accessible interface or even an indication that information is encoded within them. The generative function would only be accessible after DNA sequencing.

However, there is an inherent risk in assuming that the current technological paradigms will continue to be relevant or accessible in the distant future. DNA storage technologies could be subject to accelerated updates, transformations, and technological fashions.⁴⁵

In addition to durability concerns, the issues of interface and visibility are also critical. DNA, due to its microscopic size, inherently lacks an accessible interface for easy interpretation of the information it contains. A DNA-stored neural network cannot autonomously process or transmit its encoded information. Consequently, the vessel housing the DNA must be designed to be highly visible, as DNA alone is too small to draw attention. Effective signaling is essential to ensure that the DNA can be discovered and accessed by future generations.

Using DNA storage to preserve the AI time capsule is also complicated by DNA requiring additional layers of physical protection. While DNA can be remarkably stable under optimal conditions—such as darkness, cold, dryness, and chemical stability—its longevity is significantly reduced under less favorable conditions. For instance, DNA preserved in environments like Siberian permafrost has been successfully sequenced after 10,000 to 50,000 years.⁴⁶ However, at room temperature, DNA has a half-life of approximately 520 years. This would certainly be an insufficient lifetime to be useful for communicating warnings about radioactive waste,⁴⁷ and would significantly diminish the value of a DNA-encoded Chronoseed. Although this lifespan can be significantly extended through specialized storage conditions, the vessel containing the DNA must be carefully designed to protect it from radiation, high temperatures, and physical degradation.

The primary challenge in using DNA as a durable medium for information storage is the risk of mutation. Mutations, which are alterations in the DNA sequence, can result from environmental factors such as radiation, chemical exposure, or replication errors. Over time, these mutations can compromise the integrity of the encoded information, leading to its degradation or loss.⁴⁸

To address the potential risks of DNA destruction or mutation due to local environmental conditions, Chronoseed would be distributed in space. This arrangement acts as a backup system that preserves information even if some portions are lost or damaged over time. The challenge of durability is therefore closely tied to the question of geographic distribution. The distributed design of *Chronoseed* includes many clusters of *seeds* placed in a loosely regular pattern across different locations worldwide. While not scattered at random, its placement avoids a strictly uniform grid, ensuring that even if individual instances shift or migrate due to environmental changes, the overall structure of dispersal remains detectable and coherent. This strategic distribution increases the likelihood of Chronoseed's eventual discovery and decoding, despite long-term geological or climatic shifts.

Both the distributedness and the generativity of Chronoseed can be implemented in multiple ways, depending on the technological sophistication of the discoverers. Three distinct design proposals

⁴¹ Abdel-Rahman et al., “Modular Robotic Swarms.”

⁴² Abdel-Rahman et al., “Modular Robotic Swarms.”

⁴³ Shomorony and Heckel, “Information-Theoretic Foundations.”

⁴⁴ For some early attempts at making it computable within DNA, see Solanki et al., “Neural Network Execution.”

⁴⁵ Mendell, et al., “Matters (and Metaphors).”

⁴⁶ Poinar and Stankiewicz, “Protein Preservation,” 8426–31.

⁴⁷ Heinis, et al., “Survey of Information.”

⁴⁸ Peck, and Lauring, “Complexities of Viral Mutation.”

emerge: (1) *robust multilayered seeds with an embedded neural network*, hidden in dark, cold places for a DNA sequencing civilization, which can directly decode and reconstruct them; (2) *robust five-dimensional memory crystal disks* that can be decoded with magnifying apparatus for technologically simpler civilization; and (3) *neural network stored in viral DNA*, transmitted as a benign viral infection preserved across living hosts through generations for a civilization familiar with DNA sequencing. While DNA embedding within multilayered seeds remains the primary and most viable approach, the additional proposals, five-dimensional memory crystals and viral DNA storage, serve as speculative backups, addressing scenarios in which alternative preservation strategies might be necessary.

1. For civilizations equipped with DNA sequencing technology, Chronoseed could take the form of multilayered seeds, strategically placed in dark, cold environments. The seeds would be resistant to extreme conditions, protecting the DNA-encoded instructions necessary for building and reconstructing the neural network. The outer surface of each seed would feature inscriptions marking it as a computational object, signaling its artificial nature and suggesting engagement beyond mere observation. These markings would serve as a minimal discovery guide, prompting future discoverers to recognize its significance as something to be decoded rather than a natural or inert artifact. Chronoseed's outer shell would be flexible and suitable for small-scale manufacturing and coating—not damaging DNA during the encapsulation process. Additionally, the used material would have to be exceptionally durable, capable of lasting up to 250,000 years.

Parylene is one of the main candidates for this outer layer.⁴⁹ The subsequent layers underneath would include an anti-radiation layer to protect the DNA from mutations and a thermal insulation layer to maintain a stable, low-temperature environment.⁵⁰ At the core, the neural network's architecture, weights, and functions would be encoded into the DNA's quaternary code through a process that converts binary data into sequences of adenine (A), thymine (T), cytosine (C), and guanine (G). The DNA would be further encapsulated in protective silica, providing an additional barrier against environmental degradation.⁵¹

However, the capability to sequence DNA may not be guaranteed over long timescales. Additionally, the safest storage locations—such as permafrost regions or deep caves with consistently low temperatures and minimal humidity—are increasingly rare and may be difficult to locate or access in the future. This raises concerns about potential geopolitical and socioeconomic inequalities. Civilizations living in regions with warm temperatures might not have the opportunity to discover the seeds, effectively creating a situation where access to this preserved knowledge is unevenly distributed.

2. For civilizations lacking DNA sequencing or advanced data storage, these distributed *seeds* could be crafted as small disks from five-dimensional memory crystals—a nanostructured glass material capable of storing up to 360 terabytes in its largest form. Engineered for extreme durability, these crystals can preserve data for billions of years, enduring temperatures up to 1,000°C, withstanding impact forces up to 10 tons per square centimeter, and remaining stable even under prolonged cosmic radiation. This technology has already been applied to preservation efforts, such as the University of Southampton's project to store the entire human genome for future generations.⁵²

However, for societies with simpler technology, only the limited visible surface of the crystal may be accessible, readable with a magnifying glass. This restricts the depth and complexity of encoded information, as the time capsule's generative potential would remain out of reach.

3. Of all the proposed scenarios, embedding information within a benign virus is the most speculative, yet it presents an intriguing possibility for long-term preservation. As environments become increasingly unpredictable, using a virus as a time capsule could exploit its natural ability to integrate into host genomes and persist across generations.

This concept draws from the existence of endogenous retroviruses—viral sequences passed down for tens of thousands of years, some inherited from Neanderthals through ancient interbreeding.⁵³ The virus would be designed to remain harmless while being transmitted from one generation to the next, ensuring the preservation of critical data over millennia. DNA viruses are especially suitable for this purpose because they are stable and have relatively low mutation rates compared with RNA viruses, which are more prone to rapid degradation and frequent mutations.⁵⁴ A harmless, transmissible DNA virus could thus ensure the preservation of critical data over thousands of years, embedding information securely within human heredity and reducing the likelihood of information-altering mutations over time.

⁴⁹ Gluschke et al., "Parylene Coating System."

⁵⁰ Yu et al., "Environmental DNA Decay," 3178.

⁵¹ Grass et al., "Robust Chemical Preservation."

⁵² Baker, "Human Genome"; SPhotonix, "Pioneering the Future"; University of Southampton, "Human Genome."

⁵³ Marchi et al., "Neanderthal and Denisovan Retroviruses," R994–5; Weiss, "Human Endogenous Retroviruses."

⁵⁴ Peck and Lauring, "Complexities of Viral Mutation."

This method addresses the limitations of traditional storage mediums, which are prone to degradation or obsolescence, assuming future DNA sequencing technologies remain viable. However, it raises significant ethical concerns. Individuals may not want to be involuntary carriers of such information, further complicating the ethical landscape of this approach. While bodies already serve as living time capsules of evolutionary history, embedding outdated AI models in human DNA could infringe on bodily autonomy, preventing living beings from evolving without being anchored to a particular point from the past.

Moreover, embedding crucial information into DNA presupposes that such modifications are rare. According to Peter Watts,⁵⁵ however, pervasive biotechnological experimentation renders DNA alteration commonplace. In such a scenario, continuous genetic modifications could make it difficult to distinguish embedded information over time.

The proposed form factors do not exhaust the possibilities of embedding a neural network in tangible materials but demonstrate how deep-time communication can move beyond science fiction into a concrete, although speculative, technological proposition. At the core of Chronoseed is the idea of preserving cognitive function rather than merely archiving static information, with DNA storage emerging as a particularly promising medium for encoding a generative neural network. A new era of biocomputing, DNA storage, and hybrid organic systems is rapidly unfolding, requiring us to rethink what computation is—and where it happens. Chronoseed operates within this trajectory, leveraging DNA's unparalleled data density and longevity while remaining adaptable to future decoding methods. Recent developments, such as Biomemory's proof-of-concept DNA memory card and nanopore sequencing technologies, highlight DNA's potential as a very dense, stable storage medium, offering a plausible alternative to silicon-based systems prone to rapid obsolescence.⁵⁶ However, its long-term viability as a carrier for an interactive, self-executing system remains contingent on continued advancements in retrieval, interface design, and integration with emerging computational paradigms. As computation moves beyond traditional hardware, Chronoseed forces us to consider forms in which computational artifacts can persist in the wild, or even become part of living systems.

8 Conclusion

This paper devises a design framework for a *generative AI time capsule*, exploring how cognitive function—rather than just information—might be preserved across deep time. Through an analysis of historical and contemporary time capsules, speculative proposals in nuclear semiotics, and artifacts that have unintentionally functioned as time capsules (such as archaeological discoveries and preserved texts), we identify key misconceptions and challenges that emerge in long-term communication and knowledge transmission. These include semantic context drift, the evolution and death of languages, the paradox of technological obsolescence (where more advanced technologies tend to have shorter lifespans), the assumption of the possibility of the universal human representation, and the tendency to assume that future humanity will be civilizationaly homogeneous, rather than politically, culturally, and technologically fragmented.

To systematically address these challenges, the paper introduces a design space structured around three dimensions: (1) completeness, (2) durability, and (3) accessibility. This framework provides a means of evaluating the feasibility of any time capsule intended to remain interpretable across extended timescales, regardless of shifts in technological, linguistic, or epistemic paradigms. Within this space, the paper introduces Chronoseed as a speculative device—a generative time capsule that embeds a neural network within DNA-encoded kernels distributed across diverse environments. Unlike conventional time capsules that rely on static archiving, Chronoseed's generative function would allow for interaction, reconstruction, and engagement, adapting to the interpretative capacities of future discoverers.

That would be possible because Chronoseed would accept future inputs, allowing it to simulate contemporary cognitive functions within the epistemic contexts of its discoverers. Rather than offering a fixed repository of knowledge, it would function as both a multimodal archive and an interactive system, incorporating multiple forms of media to create a multi-sensorial Rosetta Stone effect where possible. Additionally, it would embed simplified instructions for executing its functions and constructing an interface—guidelines that future discoverers could either follow or modify, integrating their own data and methodologies to engage with its generative potential.

The problem of the interface emerges as a central challenge across the proposed material forms of Chronoseed and its potential discovery scenarios. Beyond the technological complexity of retrieval, preserving a cognitive function within DNA and enabling its execution require the assembly of a compatible interface. However, embedding such a physical interface within individual seeds risks compromising their robustness and durability. To address this, instructions for assembling an interface are encoded within the DNA, stored alongside but not within the neural network itself, providing future discoverers with the means to reconstruct a functional system without structurally compromising the preserved cognitive function.

⁵⁵ Watts, *Echopraxia*.

⁵⁶ Goldman, *Future Computing*.

DNA storage emerges as the best candidate for the material embedding of Chronoseed because of its unmatched data density and extreme longevity under optimal conditions. Unlike silicon-based storage, which is prone to technological obsolescence and environmental degradation, DNA can encode vast amounts of information in a compact, durable format while remaining decodable across diverse technological paradigms. As computation extends beyond traditional hardware, Chronoseed compels a reconsideration of how computational artifacts might endure outside controlled environments, interface with biological systems, or be reassembled independently of any fixed technological infrastructure.

Analyzing the discovery space of Chronoseed through speculative scenarios about future civilizations serves as a productive design investigation. It becomes clear that Chronoseed is not only a speculative artifact for long-term preservation but also a means of rethinking how knowledge systems themselves are structured, transmitted, and reinterpreted. By framing cultural preservation through cognitive processes—models of thought, associative logics, and sense-making structures—it foregrounds the dynamic nature of epistemic environments rather than treating knowledge systems as static cognitive architectures. This investigation moves beyond concerns of endurance and retrieval, prompting deeper inquiry into how collective thinking is shaped, stabilized, captured, and potentially reanimated in unfamiliar contexts.

Ultimately, Chronoseed operates less as a solution to the problem of deep-time preservation and more as a challenge to its usual premises. Rather than presenting a definitive design proposal, it articulates the tensions between preservation, intelligibility, and transformation, acknowledging how the emergence of synthetic intelligence reconfigures the very landscape in which these concerns unfold.

Bibliography

- Abdel-Rahman, Amira, Christopher Cameron, Benjamin Jenett, Miana Smith, and Neil Gershenfeld. "Self-Replicating Hierarchical Modular Robotic Swarms." *Nature Communications Engineering* 1 (November 2022): 35. <https://doi.org/10.1038/s44172-022-00034-3>.
- Arch Mission Foundation. "Preserving Knowledge, Forever." Accessed November 10, 2024. <https://www.archmission.org/missions>.
- Aspect, Alain, Philippe Grangier, and Gérard Roger. "Experimental Realization of Einstein–Podolsky–Rosen–Bohm Gedankenexperiment: A New Violation of Bell’s Inequalities." *Physical Review Letters* 49, no. 2 (1982): 91–94.
- Assael, Yannis, Thea Sommerschild, and Jonathan Prag. "Restoring Ancient Texts Using Deep Learning: A Case Study on Greek Epigraphy." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computing Linguistics, 2019. <https://aclanthology.org/D19-1668/>.
- Bacon, Bennett, Azadeh Khatiri, James Palmer, Tony Freeth, Paul Pettitt, and Robert Kentridge. "An Upper Palaeolithic Proto-Writing System and Phenological Calendar." *Cambridge Archaeological Journal* 33, no. 3 (2023): 371–89. <https://doi.org/10.1017/S0959774322000415>.
- Baker, Harry. "Human Genome Stored Inside Near-Indestructible '5D Memory Crystal' That Could Survive to the End of the Universe," *Livescience.com*, September 25, 2024. <https://www.livescience.com/technology/human-genome-memory-crystal>.
- Biersdorfer, J. D. "A Time Capsule of Human Creativity, Stored in the Sky." *New York Times*, July 27, 2023. <https://www.nytimes.com/2023/07/27/arts/design/lunar-codex-time-capsule-moon.html>.
- Bratton, Benjamin. *Cognitive Infrastructures: Synthetic Intelligence in the Wild*. Lecture at Central Saint Martins, London, July 3, 2024. <https://www.e-flux.com/announcements/591674/cognitive-infrastructures-synthetic-intelligence-in-the-wild/>.
- Can, Muhammed. "Under the Leadership of Our President: 'Potemkin AI' and the Turkish Approach to Artificial Intelligence." *Third World Quarterly* 44, no. 2 (2023): 356–76. <https://doi.org/10.1080/01436597.2022.2147059>.
- Chandler, David. "Flocks of Assembler Robots Show Potential for Making Larger Structures." *MIT News* (2022). <https://news.mit.edu/2022/assembler-robots-structures-voxels-1122>.
- Chapman, Kit. "Speaking to the Future." Science History Institute. Accessed August 30, 2024. <https://www.sciencehistory.org/stories/magazine/speaking-to-the-future/>.
- Dong, Haiwei, and Shuang Xie. "Large Language Models (LLMs): Deployment, Tokenomics and Sustainability." Preprint, arXiv, May 27, 2024. <https://arxiv.org/abs/2405.17147v1>.
- Fabbri, Paolo, and Françoise Bastide. "Living Detectors and Complementary Signs: Cats, Eyes, and Sirens." *Linguistic Frontiers* 5, no. 3 (2022): 10–13. <https://doi.org/10.2478/lf-2022-0008>.
- Foucault, Michel. *The Archaeology of Knowledge*. Routledge, 2002.
- Gluschke, Jan Göran, Felix Richter, and Adam P. Micolich. "A Parylene Coating System Specifically Designed for Producing Ultra-Thin Films for Nanoscale Device Applications." *Review of Scientific Instruments* 90, no. 8 (2019): 083901. <https://doi.org/10.1063/1.5099293>.
- Golding, Johnny, Martin Reinhart, and Mattia Paganelli. *Data Loam: Sometimes Hard, Usually Soft. The Future of Knowledge Systems*. De Gruyter, 2020.
- Goldman, Nick. "Future Computing: DNA Hard Drives. Kurzgesagt—In a Nutshell. World Economic Forum. Posted March 10, 2015, by World Economic Forum. YouTube, 9:51. <https://www.youtube.com/watch?v=tBvd7OSDGgQ>.

- Grass, Robert N., Reinhard Heckel, Michela Puddu, Daniela Paunescu and Wendelin Stark. "Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes." *Angewandte Chemie International Edition* 54, no. 8 (2015): 2552–55. <https://doi.org/10.1002/anie.201411378>.
- Hawranek, Maria. "Stanisław Lem & Messages to the Future . . . Through Plant DNA?," translated by Anna Potoczny. *Culture.pl*. Last updated October 21, 2021. <https://culture.pl/en/article/stanislaw-lem-messages-to-the-future-through-plant-dna>.
- Heinis, Thomas, Roman Sokolovskii, and Jamie J. Alnasir. "Survey of Information Encoding Techniques for DNA." *ACM Computing Surveys* 56, no. 4 (April 2024): 107. <https://doi.org/10.1145/3626233>.
- Herzog, Lisa. "What's Wrong with the 'Marketplace of Ideas'?" In *Citizen Knowledge: Markets, Experts, and the Infrastructure of Democracy*, edited by Lisa Herzog. Oxford University Press, 2023. <https://doi.org/10.1093/oso/9780197681718.003.0005>.
- Ialenti, Vincent. *Deep Time Reckoning: How Future Thinking Can Help Earth Now*. The MIT Press, 2020.
- International Atomic Energy Agency. *Radiation Protection and Safety of Radiation Sources: International Basic Safety Standards*. International Atomic Energy Agency, 2014. https://www-pub.iaea.org/mtcd/publications/pdf/pub1419_web.pdf.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, et al. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55, no. 12 (2023): 248. <https://doi.org/10.1145/3571730>.
- Jo, Eun Seo, and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." In *FAT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2020. <https://doi.org/10.1145/3351095.3372829>.
- Lindgren, Kristian. "Information Theory for Complex Systems." *Lecture notes* (January 2003, updated in 2014). Department of Physical Resource Theory, Chalmers and Göteborg University, 2003.
- Lowenthal, David. *The Past Is a Foreign Country*. Cambridge University Press, 1985.
- Lyotard, Jean-François. *The Postmodern Condition: A Report on Knowledge*. Manchester University Press, 1984.
- Mansoux, Aymeric, Brendan Howell, Dušan Barok, and Ville-Matias Heikkilä. "Permacomputing Aesthetics: Potential and Limits of Constraints in Computational Art, Design and Culture." *Ninth Computing Within Limits*. LIMITS, 2023. <https://doi.org/10.21428/bf6fb269.6690fc2e>.
- Marchi, Emanuele, Alex Kanapin, Matthew Byott, Gkikas Magiorkinis, and Robert Belshaw. "Neanderthal and Denisovan Retroviruses in Modern Humans." *Current Biology* 23, no. 22 (November 2013): R994–5. <https://doi.org/10.1016/j.cub.2013.10.028>.
- Mazzucchelli, Francesco, and Nanta Novello Paglianti. "How to Remember a Place to Forget? The Semiotic Design of Deep Geological Nuclear Repositories, from Long-Term Communication to Memory Transmission." *Linguistic Frontiers* 5, no. 3 (December 2022): 22–36. <https://doi.org/10.2478/lf-2022-0026>.
- Memory of Mankind. "MoM." Accessed September 10, 2024. <https://www.memory-of-mankind.com/>.
- Mendell, Madeleine, Mél Hogan, and Deb Verhoeven. "Matters (and Metaphors) of Life and Death: How DNA Storage Doubles Back on Its Promise to the World." *Canadian Geographies/Géographies Canadiennes* 66, no. 1 (March 2022): 37–47. <https://doi.org/10.1111/cag.12741>.
- Moynihan, Thomas. "If a Victorian Historian Had His Way, There'd Be a Giant Time Capsule Under Stonehenge." *BBC Future*, July 9, 2024. <https://www.bbc.com/future/article/20240709-if-a-victorian-historian-had-his-way-thered-be-a-giant-time-capsule-under-stonehenge>.

- Navigli, Roberto, Simone Conia, and Björn Ross. "Biases in Large Language Models: Origins, Inventory, and Discussion." *Journal of Data and Information Quality* 15, no. 2 (2023): 10. <https://doi.org/10.1145/3597307>.
- Parkinson, Richard B., Whitfield Diffie, Mary Fischer, and R. S. Simpson. *Cracking Codes: The Rosetta Stone and Decipherment*. University of California Press, 1999.
- Peck, Kayla M., and Adam S. Luring. "Complexities of Viral Mutation Rates." *Journal of Virology* 92, no. 14 (2018): e01031-17. <https://doi.org/10.1128/jvi.01031-17>.
- Pierron, Robin, Sylvain Leclar, Julien Zelgowski, Pierre Pfeiffer, Frédéric Mermet, and Joël Fontaine. "Etching of Semiconductors and Metals by the Photonic Jet with Shaped Optical Fiber Tips." *Applied Surface Science* 418 (2017): 452–55. <https://doi.org/10.1016/j.apsusc.2017.01.277>.
- Poinar, Hendrik N., and B. Artur Stankiewicz. "Protein Preservation and DNA Retrieval from Ancient Tissues." *Proceedings of the National Academy of Sciences* 96, no. 15 (1999): 8426–31.
- Raffel, Colin, Noam Shazeer, Adam Roberts, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21, no. 140 (2020): 1–67.
- Rao, Rajesh P. N. "The Indus Script and Economics: A Role for Indus Seals and Tablets in Rationing and Administration of Labor." Unpublished, n.d.
- Rao, Venkatesh. "Superhistory, Not Superintelligence." Ribbonfarm Studio. May 11, 2021. <https://studio.ribbonfarm.com/p/superhistory-not-superintelligence>.
- Renganathan, Vasu. "Tracing the Trajectory of Linguistic Changes in Tamil: Mining the Corpus of Tamil Texts." *IJDL International Journal of Dravidian Linguistics* 43 (2013): 351–65.
- Sanusi, Tife. "Reviving Lost Tongues: How AI Battles Language Extinction." Deepgram, August 17, 2023. <https://deepgram.com/learn/language-ai-cultural-preservation>.
- Schalansky, Judith. *An Inventory of Losses*. MacLehose Press, 2021.
- Schmandt-Besserat, Denise. "The Evolution of Writing." In *International Encyclopedia of the Social & Behavioral Sciences*, edited by James D. Wright. Elsevier, 2015. <https://sites.utexas.edu/dsb/tokens/the-evolution-of-writing/>
- Shomorony, Ilan, and Reinhard Heckel. "Information-Theoretic Foundations of DNA Data Storage." *Foundations and Trends® in Communications and Information Theory* 19, no. 1 (2022): 1–106. <https://doi.org/10.1561/01000000117>.
- Solanki, Arnav, Zak Griffin, Purab Ranjan Sutradhar, et al. "Neural Network Execution Using Nicked DNA and Microfluidics." *PLOS One*, 18, no. 10 (2023): e0292228. <https://doi.org/10.1371/journal.pone.0292228>.
- Sphotonix. "Pioneering the Future of Data Storage and Optical Innovation." Accessed November 10, 2024. <https://sphotonix.com/>.
- Strachan, Tavares. *The Encyclopedia of Invisibility*. Marian Goodman Gallery, 2021. <https://www.mariangoodman.com/artists/312-tavares-strachan/works/55052/>.
- Tishby, Naftali, and Noga Zaslavsky. "Deep Learning and the Information Bottleneck Principle." *2015 IEEE Information Theory Workshop (ITW)*, Jerusalem, 2015. <https://doi.org/10.1109/ITW.2015.7133169>.
- Tolstaya, Tatyana. *The Slynx*, translated by Jamey Gambrell. NYRB Classics, 2007.
- Trauth, Kathleen M., Stephen C. Hora, and Robert V. Guzowski. "Expert Judgment on Markers to Deter Inadvertent Human Intrusion into the Waste Isolation Pilot Plant." Sandia Report SAND92-1382 UC-721. Sandia National Laboratories, November 1993. <https://gwern.net/doc/technology/1993-trauth.pdf>.

- University of Southampton. "Human Genome Stored on 'Everlasting' Memory Crystal." *News*, September 2024.
<https://www.southampton.ac.uk/news/2024/09/human-genome-stored-on-everlasting-memory-crystal-page>.
- US Department of Energy. "Reducing the Likelihood of Future Human Activities That Could Affect Geologic High-Level Waste Repositories." Technical Report, United States Department of Energy, May 1984. <https://doi.org/10.2172/6799619>.
- US Nuclear Regulatory Commission. "Backgrounder on Plutonium." Accessed August 25, 2024.
<https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/plutonium.html>.
- Watts, Peter. *Echopraxia*. Head of Zeus, 2014.
- Weiss, Robin A. "Human Endogenous Retroviruses: Friend or Foe?" *APMIS*, 124, no. 1–2 (2016): 4–10.
<https://doi.org/10.1111/apm.12476>.
- Woods, Christopher, ed., with Emily Teeter and Geoff Emberling. *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. OIMP no. 32. Oriental Institute of the University of Chicago, 2010.
- Yablon, Nick. *Remembrance of Things Present: The Invention of the Time Capsule*. University of Chicago Press, 2019.
- Yu, Xue, Jiaying Zhou, Jun Wei, and Xueqiang Lu. "Temperature May Play a More Important Role in Environmental DNA Decay than Ultraviolet Radiation." *Water* 14, no. 19 (2022): 3178.
<https://doi.org/10.3390/w14193178>.
- "Zeitschrift für Semiotik: Band 6, Heft 3" [*Journal of Semiotics*]. *Technische Universität Berlin*, 1984. Accessed November 8, 2020.
https://web.archive.org/web/20201108010948/https://www.semiotik.tu-berlin.de/menue/zeitschrift_fuer_semiotik/zs_hefte/bd_6_hft_3/#c185967.



The Chronoceptual Governor

Michelle Chang

University of
Washington

Tyler Farghly

University of
Oxford

Yannis Siglidis

Ecole Des Ponts
ParisTech

Abstract

Traditionally, philosophy views the temporal relationship between technology and life as one where an unnatural clock of the first disrupts and alienates the natural clock of the second. However, when viewed as an ecological force, technology's relationship with time can be reframed inside a context of adaptation. This paper lays out a framework for understanding the ecology of temporal relationships between actors and their environments via two different notions of time: (1) chronoception (the way an actor perceives time), and (2) characteristic time (the amount of change that appears in an environment). In the latter, computation acts as the means through which actors can expand their chronoception to further timescales of characteristic time. This expansion, however, is not neutral. Through their computational expansion, actors do not merely perceive but also affect the characteristic time of their environments, which are constructed by nested and interwoven temporalities, creating an evolutionary force of temporal misalignment. This framework enables us to rethink cybernetic governance, not as a problem of control but as a problem of temporal mediation, where computation serves not to impose its own temporality but to mediate and coordinate actors across entangled temporalities.

Keywords

time; chronoception; computation; ecology; technology; adaptation.

1 Introduction

Typically, when technology is to be reconciled with time, it is through measurement. Time becomes an axis that technology either refines or occupies. Technology's time is often thought of as constant—an objective measure that necessitates calibration if it should diverge from perfection. One modernist dream was to define and democratize absolute time. This was also the Fordist dream—in which the pocket watch became an essential accessory for Tronti's social factory. Implemented within the gears of mechanical timepieces and subsequently manifested for the age of computation as standardized systems, like processor time or UNIX time, the now materialized Fordist clock enabled not just the synchronization of individual actions but also of society.¹ *Clock speed*, a fundamental measure of a computer processor's performance, is now the fastest rate at which matter-as-information can be parsed by a single CPU. Moore's Law famously predicts that the number of transistors on a microchip will double every two years. Indeed, MOSFET sizing has downsized from 6 μm in 1974 to roughly 2 nm in 2024. In parallel, clock speed has increased from 2 MHz to 6.2 GHz.²

However, the idea that clock speed is a single objective value is an oversimplification. Rather, its *phenomenal time*—the subjective, felt experience of change encountered by those who engage with it—arrives to us as the imbrication of many layers of hardware, software, and user interface and, in turn, varies according to our familiarity and our time-sensitive perceptual abilities.³ This is crucial, as the *time* in which humans perceive technology is often the time in which they can use it. However, modern digital interfaces tend to push users to the limits of how fast they can use them.⁴ Seen in their totality, improvements in distributed computing, cloud infrastructure, and handheld personal devices created a complex multilayered information infrastructure that has managed to adapt and evolve our app-driven general intellect to a fully automated, high-speed, post-broadcast-era consumption culture, where ordinary, everyday moments are constantly mediated by an almost metabolic, one-minute-reel attention economy. In short, the lithosphere was transformed into a “racing experience.”⁵

Accordingly, when engaging with the subject of time and technology, philosophy historically focuses on how technology's clock disrupts and accelerates our phenomenal experience of time. Irrespective of how many cycles of design and adaptation it goes through, technology always seems to succeed in making humanity tick according to its whims.⁶ Inherently, this is what Marx tried to argue through the concept of metabolic rift, which posits that technology's reproduction time does not align with or disturbs that of nature.⁷ Various downstream theories further view technology as an uninhabitable temporal *environment*. Grossly, they could be grouped among those of *acceleration*, argued by Heidegger, Simmel, and Carey,⁸ or those of *simultaneity*, discussed by Castells, Virilio, and McLuhan.⁹ Acceleration reduces technology's influence on time to one of increase in speed and efficiency, whereas simultaneity sees how (especially digital) technologies change the speed of certain operations by orders of magnitude (for example, that of instant messaging), which leads to a different notion of physical space.

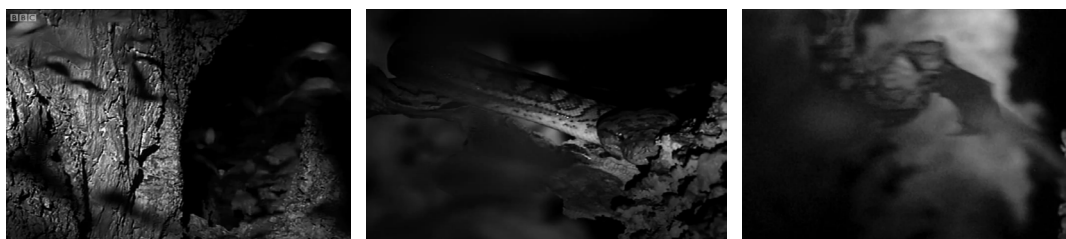


Figure 1 A snake's local temporal adaptation. Inside a cave, a bottleneck is being formed where bats group to pass (left). A snake discovers that point and sneaks to a convenient spot, waiting to snatch its prey (middle). Due to the bottleneck, the bats move much more slowly than they would individually. Their characteristic time now aligns with the snake's chronoception, which works to the snake's advantage (right; BBC Earth, “Snakes Hunt Bats”).

¹ Thompson, “Time, Work-Discipline.”

² Wikipedia, “(2024, August 20). *Clock rate*,”. In *The Free Encyclopedia*. Last modified March 29, 2025, 00:46 (UTC), https://en.wikipedia.org/wiki/Clock_rate.

³ Nielsen, “Time Scales of UX.”

⁴ Kuijer and Laschke, “Post-Growth Society.”

⁵ Tan and Hu, “Speed and Passion.”

⁶ Colvile, “Great Acceleration.”

⁷ Marx, *Capital*.

⁸ Heidegger, *Being and Time*; Staudacher, “Simmel's Sociology of Time”; Carey, *Communication as Culture*.

⁹ Castells, *Space of Flows*; Virilio, “Illusions of Zero Time”; McLuhan, *Understanding Media*.

1.1 Definitions

If life carries an essential clock, it is mainly dictated by an organism's metabolic rate, which describes the general rate of energy consumption needed to properly function. It emerges from a set of physical, chemical, biological, behavioral, and computational evolutionary traits that fuel an animal's adaptation to its constantly evolving environment. At the same time, an animal can effectively bind all these traits into one coherent experience of subjective time, which we refer to as *chronoception*. Chronoception is a subjective quality that describes how quickly an autonomous individual, or *actor*, perceives information in a unit of time. In contrast, the set of organic and biotic matter relevant to an actor's survival, with which it must engage daily, is what we refer to as its *environment*. To define the rate of change in the environment, we introduce the notion of *characteristic time* from dynamical systems. Characteristic time measures the timescale over which certain systems develop. For example, the same physical operation of diffusion occurs almost instantaneously in the material substrate of electronics, whereas it is comparatively slow when it takes place inside biological cells. This conception of time echoes Eddington's entropy clock, which frames time through the entropic evolution of a physical system.¹⁰ This measure is objective in nature, yet its dynamics can be complex, evolving, and recursive. Between chronoception and characteristic time stands *computation*, which refers to the ability of an actor to model, anticipate, and plan its environment.

We can observe these elements in play in the context of animals surviving in the wild. Generally, organisms tend to situate themselves within ecosystems that strike a balance between their chronoceptual abilities, the characteristic time of the environment, and the organism's capacity of computation. There are often trade-offs to be made here: due to energy and mass constraints, for instance, large animals tend to have slower chronoception than smaller animals,¹¹ and their lifespan tends to be larger, while their moving speed tends to be lower.¹² An interesting case that demonstrates the endless flux and specificity of temporal adaptation can be found in a tale of snakes and bats. Caves, the habitat of bats, change relatively slowly, appearing effectively static for an animal like a bat, thus serving as a local optimum for roosting. These same caves also attract snakes, which eagerly explore them in search of prey, using their thermal sensing to their advantage. The snake has adapted to occupy two neighboring temporal "strata." In one, it is a slow-moving entity looking to convince its prey that it is part of the surrounding static (and thus safe) environment. In the other, it is a dynamic entity, attacking at a speed faster than its prey can react. However, a single bat moves too quickly for a snake to reliably snatch, even if it flies within striking range. Yet, the snake employs an interesting strategy by locating a bottleneck formed inside the cave, wherein, even if bats fly at extreme speeds, their movement as a group decreases their individual speeds and their density increases significantly. Exploiting this phenomenon, the snake throws its jaws toward a slowly moving sea of life (see Figure 1).

1.1 Structure

In this essay, we will try to reframe the relationship of time to technology from one that is opposed to a disturbing, self-directed, and alienated clock to one that can be thought of in ecological terms. To clarify this further, first in section 2, we will outline a set of temporal zones that exists between an actor and its environment. Then, in section 3, we show how computation can affect these zones. In section 4, we will see what is behind the temporality of an actor and analyze the effect of temporal interactions. Finally, in section 5, we discuss the political implications of our framework, highlighting how time reframes the role of technology as a governor in the context of cybernetics.

2 Chronoconceptual Learnability

The subjective experience of time has long been a subject of discussion within psychology, where variables such as age, emotional state, and the influence of psychoactive substances have been linked to variations in chronoception.¹³ Notably, Henri Bergson focused on the perception and psychological experience of duration, differentiating between the subjective notion of duration and the objective temporality governed by physical processes—what we refer to in this paper as *characteristic time*.¹⁴ Foundational to Bergson's notion of duration was the understanding that time is experienced through the recognition and collection of change: through memory, which in modern terms would be the key frames extracted through pattern recognition and modeling of an actor's environment. Hence, the capacity to perceive change at a particular rate is intimately connected to the speed at which time can be sensed—the speed of chronoception.

In this light, chronoception is not merely an abstract qualitative phenomenon but something that can be quantified and measured empirically. The perception of change and the instantiation of change are two sides of the same coin: one occurs within the actor and one outside, within the environment. With our formulation of chronoceptual and characteristic time through the detection and creation of change, we have established axes along which subjective experience and objective reality intersect.

¹⁰ Eddington, *Physical World*.

¹¹ Healy et al., "Metabolic Rate."

¹² West, *Scale*.

¹³ Aday et al., "Psychedelic Drugs and Perception" 2021; Gable et al., "How Does Emotion Influence." 2022

¹⁴ Bergson, *Time and Free Will*.

Here, the crucial question arises: what transpires when these temporalities diverge? What is experienced when the speed of chronoception doesn't align with that of characteristic time? In Figure 2, we map out the two-dimensional space consisting of different combinations of chronoceptual and characteristic speed and identify phenomena that appear in five distinct regions: time-averaging, aliasing, learnability, myopia, and timescale separation.

The first region, *time-averaging*, emerges when the characteristic speed of the environment far exceeds the chronoceptual speed of the actor. In this scenario, which characterizes how we perceive, for example, quantum mechanical phenomena such as electricity, the subject cannot but perceive the environment as a time-averaged blur, where the nuances of changes are lost in a near-amorphous projection of reality. At the opposite extreme lies *timescale separation*, an example of which is geological time, where the characteristic speed of the environment is so slow relative to the actor's chronoception that it appears completely static and devoid of life, even if the environment does undergo change. In this context, the actor perceives the environment as unchanging, making any form of exchange or responsive interaction between actor and environment impossible. Drawing analogies from photography, time-averaging is akin to viewing the environment through the lens of long-exposure photography, where rapid movements are transformed into streaks of light, erasing the fine details of motion and form. Conversely, timescale separation is analogous to fast-shutter-speed photography, which captures a dynamic environment in such a way that it appears frozen and lifeless, rendering the scene unusually inert. These two extremes of temporal ratios reveal the presence of a temporal filter, a mechanism that regulates the interaction between characteristic and chronoceptual speeds and in this way prevents radical mismatches from interfacing.

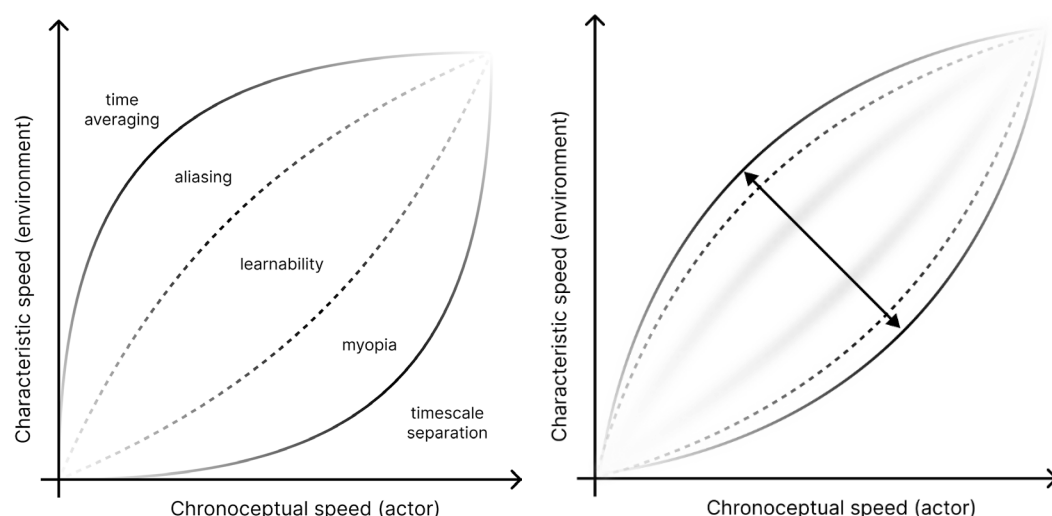


Figure 2 Relative temporal relationships between actor and environment and their computational expansion. On the left, we display, as a constant ratio of chronoceptual and characteristic speeds, different regions of phenomena of temporal misalignment, at the center of which is learnability, the most relevant zone to temporal adaptation. On the right we depict how computation expands the zone of learnability to further, previously inaccessible zones.

Yet, there are subtler mismatches in chronoceptual ratios, where perception remains possible but is fundamentally distorted. *Aliasing* describes the phenomena that occur when the characteristic speed of an environment slightly outpaces the chronoceptual speed of the actor. In this case, the actor is capable of only partially capturing the change occurring in the environment and is thus prone to misinterpreting the nature of the environment's evolution. This effect is exploited in the creation of motion pictures, where rapidly flickering images are mistaken for continuous motion. Conversely, when the characteristic speed lags just behind the actor's chronoception, we enter the region of *myopia*. Here, the actor becomes fixated on immediate, surface-level changes, mistaking them for the entirety of the environment's developmental trajectory. The broader, slower patterns that define the environment's true nature are obscured.

In both cases, these minor temporal mismatches distort the actor's interface with reality, revealing that accurate perception requires a certain alignment between chronoceptual speed and characteristic speed. This forms the region of *learnability*, where the actor is neither overwhelmed by rapid change nor lulled into complacency by slow dynamics. Instead, the actor is situated in a sweet spot, enabling the accurate perception of the environment. The evidence that chronoception is related to metabolic rate suggests that, as expected, evolution selects for this zone.¹⁵ On a larger scale, mathematical support for the Red Queen hypothesis, that is, the conjecture of a constantly changing state of nature, points to a sweet spot of learnability between a very fast system of mutations at the level of

¹⁵ Healy et al., "Metabolic Rate."

the actor and a comparatively slow, changing environment.¹⁶ Sociologist Hartmut Rosa even saw the human social subject as analogous to a Red Queen, as it evolves inside an accelerated, technologically advanced modern society.¹⁷

3 Computation as Chronoconceptual Expansion

Over the last half century, computation has become a planetary exoskeleton as well as a sociobiological one.¹⁸ Mirroring the development of language, computation—as unveiled through scientific instruments of climate change or mass-culture, data-driven autophagy—managed to augment, operationalize, and coordinate human cognitive faculties. In fact, computation became equivalent to an operational language of material processes across and beyond culture. By mastering robustness through digitization and optimization, computation evolved scaling and sensing to capture, analyze, and simulate phenomena well beyond natural perceptual abilities. Computation became what enabled humans to perceive themselves, equivalently, from the timescale of the universe and the timescale of the atom. When traditional views highlight how computation accelerates time, they obscure how it reveals and mediates it. Aside from defining, standardizing, and operationalizing time, computation also makes the temporality of one process apparent to another, making it possible for an actor to act on previously unexplorable dynamics.

In short, computation doesn't only disturb our chronoception, it expands it. For example, the ability to capture and manipulate data enables us to perceive and understand phenomena that were previously obscured by our chronoconceptual filter. High-speed photography, for instance, has revealed aspects of reality that were previously inaccessible, concealed within the temporal aliasing artifacts of our visual perception. The early applications of this technology served to deepen our understanding of nature, as exemplified by Eadweard Muybridge's photographic sequences of a horse in motion, which unveiled the mechanics of galloping, Ernst Mach's studies of supersonic motion, and Jean Comandon's film revealing the "growth of plants."¹⁹ Furthermore, capturing data can serve to expand our chronoception to incorporate slow characteristic speeds. Through compression, time-lapse photography extended temporal processes into digestible sequences, enabling us to observe the fluid dynamics of natural phenomena that would otherwise remain invisible to human perception.

The implications of data capture can be extended further when combined with the practice of simulation.²⁰ The origins of simulations lie in the need to study the evolution of dynamical systems. An example of a dynamical system is the trajectory of a mass particle under Newton's second law, which makes it possible to obtain a closed-form solution of that particle given properly defined initial conditions. However, for most observed systems, such closed-form solutions don't exist, making it impossible to understand their dynamics. Using the differential equations or the step function that describes how a system evolves and is often easier to retrieve or design, one can instead simulate the trajectories of the system and analyze its predictions. In this way, simulations enable us to compress time in two different ways. One is empirical, as by running a simulation faster we can both make observations and perceive changes by bringing a system's characteristic time to ours. The other is anticipatory, as an actor can prepare itself for the future by effectively compressing time.

However, the blind spot of simulations is that they often rely on approximations, as computing or modeling the dynamics of a system with precision can be intractable. This is important, as in certain areas, simulations serve as the singular potent epistemic means. In climate science, historical data acquisitions are complemented by their extrapolated trajectories of metrics of climate evolution decades into the future, offering a glimpse of long-term counterfactuals that far exceed the lifetime of any human being. This creates a form of structural uncertainty that is factored into the timescales that computation operates in and models. However, while a model inherently suffers from both epistemic and structural uncertainty, approximate answers to certain questions may still be enough to plan future action, as is for example the case in the context of climate science.

At the particle scale, simulations similarly rely on scientific models and fragmented data to stretch human temporal comprehension and perceive phenomena that evolve at the timescale of the nucleus. Here, by slowing down or speeding up phenomena not perceivable before, computation can adjust the observed characteristic time of certain processes, effectively expanding the zone of chronoconceptual learnability. As a result of its influence on perception, simulation informs action for the purposes of survival and adaptation. The expansion of the zone of chronoconceptual learnability leads to the possibility of temporal adaptation, which can come in two ways. One is the observation of a phenomenon that affects an actor, despite the actor previously being unable to perceive it. The other is through the discovery of a phenomenon taking place in a chronoconceptual zone, previously unexplored, which, after being discovered, can be exploited to the actor's reproductive advantage. However, as the actor inhabits new zones of temporality, it also has the ability to influence and, eventually, disturb them.

¹⁶ Wortel et al., "Continual Evolution."

¹⁷ Rosa, *Social Acceleration*.

¹⁸ Bratton, *The Stack*.

¹⁹ Muybridge, "Horse in Motion"; Mach and Salcher, "Photographische Fixirung"; Comandon, *La croissance des végétaux*.

²⁰ Winsberg, *Computer Simulations in Science*.



Figure 3 Actors as systems of systems: Actors are nestings of systems and of their temporalities. Source: Wang et al., “Generative Powers of Ten.”

Until now, this work has engaged with *chronoceptual speed* as the rate at which an actor perceives information in a unit of time, and with *characteristic time* as the rate at which a certain system develops or changes in a single unit of time. However, actors are multitudes of various temporalities nested and bound up within a single system—an assemblage that perceives time at a rate different from that of its constituent parts (see Figure 3). To conceptualize these systems, a handy concept that relates to the study of nature and that extends to our everyday lives is that of the *holobiont*. The holobiont is the collective whole represented by an “animal or plant with all of its associated microorganisms.”²¹ Cows, for instance, rely on bacteria in their gut to digest the grass they eat. Similarly, the Hawaiian bobtail squid relies on *Vibrio fischeri* to luminesce in the deep ocean and hide from predators.²² Even humans themselves are holobionts: we rely on a wide variety of symbiotic relationships with microorganisms living within and on us to function properly. Indeed, the human gastrointestinal tract is home to many gut microbiota that affect immune health, prevent disease, and even affect our moods.²³ Without them, we could not survive. As Gilbert states: “We have about 160 major species of bacteria in our bodies, and they all form complex ecosystems. Human bodies are and contain a plurality of ecosystems.”²⁴ Thus, humans—or any actor, for that matter—are not ontological conclusions. They, in turn, are also part of other holobionts, active constituents of other environments. In other words, actors are not just *nestings*, but are also *nested*. Depending on the point of reference used for analysis, actors can be said to be either environments unto themselves or part of another actor’s environment. All roles are relational.

We have discussed how to delineate between actor and environment. Yet, how does one account for the temporal gap between the two? Consider that when many actors are assembled into a meta-actor (an environment), the chronoception of the whole system may (and, indeed, most likely will) diverge from the individual characteristic times that make up the system. To account for this, computation can be used to bridge that gap and isolate certain temporalities from (or adjust them to) an actor’s chronoception. However, to understand the influence actors have on their environments, we need a framework that conceptualizes temporal interactions as those that occur not between clocks, but between assemblies of clocks. That is, as a constant push and pull between *temporal profiles*.

To think of the “space” of characteristic time is, colloquially, to think of a stack of temporal strata operating in parallel—as if the universe were a geological cake of disconnected layers, each moving through the world at a separate characteristic speed. All technological systems have a distinct *temporal holobiont*, which is also manifested in the context of a *stack*.²⁵ A common technology like Google Maps, for instance, could be said to enfold actors such as the human user, a mobile phone, a satellite network, and a GPS receiver. Each component deals with time at its own pace. The GPS receiver can capture a signal from a satellite faster than humans can perceive a change in position, and even faster than any orbital interference introduced by the moon (Figure 4).

²¹ Zilber Rosenberg and Rosenberg, “Role of Microorganisms.”

²² McFall-Ngai, “Noticing Microbial Worlds.”

²³ Appleton, “Gut-Brain Axis.”

²⁴ Gilbert, “Holobiont by Birth,” 75.

²⁵ Bratton, *The Stack*.

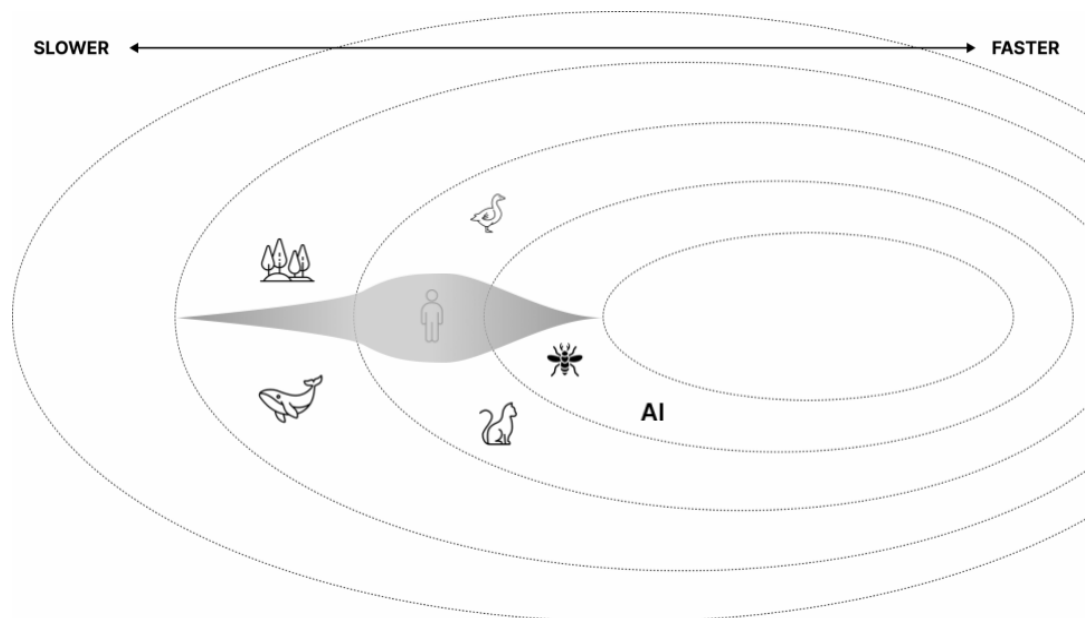


Figure 4 Nested actor: Actors are nestings of temporalities, often across temporal strata.

However, thinking of the temporal holobiont as a disconnected model does not account for interactions that occur at the boundaries of characteristic times. Such interactions inevitably arise when various characteristic times cross-contaminate and intertwine (see section 2 for examples). Taking into account the nested and interconnected structure of multiple temporalities, a temporal profile of characteristic times may offer a more helpful representation (see Figure 5). Although we utilize temporal profiles mostly as a conceptual image, we borrow them from the frequency spectrum visualization technique often used to describe the temporal variation of certain signal readouts, such as sound, voltage, or dynamical systems.²⁶ This representation is often valuable for visualizing the frequency composition of a signal across a single time frame. This kind of framework enables us to illustrate the idea that the temporal profiles of both actors and environments are not constants but the accumulation of multiple locations along the stack of temporal strata. Although the core of an actor's temporal profile might center around a certain range of characteristic times, each will nonetheless contain those of its constituent and peripheral elements, possibly tapering out at either end of the strata around and within which it is nested.

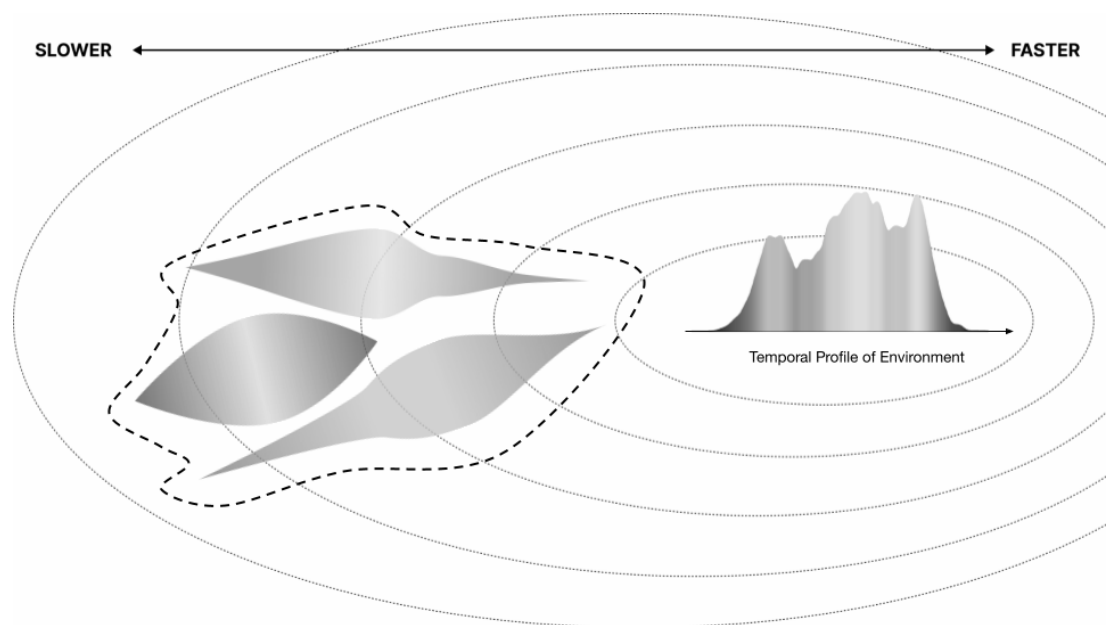


Figure 5 Temporal profiles of nested actors. Environments nest the temporalities of their constituent actors. Here, the environment's temporal profile is shown to be the aggregate of the profiles of its actors. External actors interacting with such an environment engage with the pictured temporal profiles at the regions where both the actor's and the environment's temporal profiles intersect.

²⁶ Wikipedia, "(2024, August 20). *Fourier Analysis*,". In last modified February 13, 2025, 10.48 (UTC), *The Free Encyclopedia*. https://en.wikipedia.org/wiki/Fourier_analysis

By exposing one temporal profile to another, patterns of interference emerge. We are now able to conceptualize the interaction between actors and their environments as the intersection between the profiles of characteristic time that a system and its environment exhibit at the plane of contact (see Figure 6). If both actor and environment have already coevolved to a state of equilibrium, these interactions will be fairly stable and routine. Otherwise, these areas of diffraction may result in less controllable dynamics that necessitate evolution—either of the actor or of the environment.

One example of an environment adapting to an actor might be how digital technologies have increased “the resource intensity, and therefore ecological pressures of everyday life.”²⁷ Using the washing machine as a representative example, Kuijer and Laschke note that such innovations make it possible to delegate human actions, leading to “changes in human behaviors, skills, and time management” and the subsequent accumulation of more technological gadgets.²⁸ This process of accumulation and integration is accelerating further, because the pace at which, for example, washing machines evolve in relation to humans is faster. The result is a compounding of technologies in everyday tasks and a speeding up of daily life. Individual technological inventions are setting off a flywheel in which human ecosystems and economies must continually adapt to the compounding effects of these actors. Inversely, an actor also has the ability to disrupt its environment by triggering a sequence of temporal disruptions that can disturb its underlying physical processes. A particularly representative example is the introduction of a goat population to an island, which leads to overgrazing and, subsequently, desertification.²⁹ However, on a larger scale, similar perturbations can be seen in more complex phenomena, such as the disruption of the water cycle in the context of global warming.³⁰

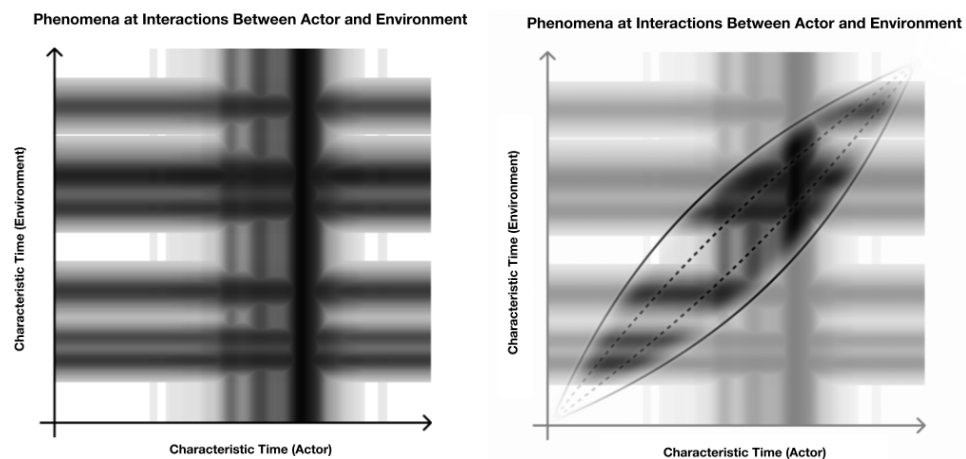


Figure 6 Temporal intersection of temporal profiles between actors and environments. On the left, interference patterns that come from the interaction between actor and environment. On the right, only some of those modes of interaction occupy a zone of learnability, thus influencing adaptation.

At a glance, we observe what seems like an almost unitary dynamic: either the actor has to adjust to the temporality of its environment or, inversely, the environment has to speed up or slow down to match the characteristic speed of the actor. This condition of temporal misalignment can be referred to as *lag*, which can be characterized as a cross-temporal gradient that guides the direction of temporal adaptation.

However, to perform such an adaptation, a system needs to be aware of how its action might influence the temporality of another system. This becomes a cybernetic problem of a system learning how to balance the dynamics of another system toward homeostasis.³¹ This reveals that in addition to what we discussed in section 3, the role of computation isn’t only to expand the chronoception of individual actors, but also to become that which mediates and aligns systems across their characteristic times. In the next section, we will discuss how this shift in perspective from the temporality of the actor to the temporality of the ecosystem, and of computation from the governor to the medium, reframes cybernetics from a project of control to one of coordination.

5 Chronoconceptual Governance

The problem of what nature might be, returns from exile among the hippies.
For a long time, it seemed like a critical gesture to insist that reality is socially constructed.
Now it seems timely to insist that the social is reality constructed.
—McKenzie Wark, *Capital is Dead*

²⁷ Kuijer and Laschke, “Post-Growth Society,” 2.

²⁸ Kuijer and Laschke, “Post-Growth Society,” 2.

²⁹ Arianoutsou-Faraggitaki, “Desertification by Overgrazing.”

³⁰ Intergovernmental Panel on Climate Change, *Climate Change* 2023.

³¹ Wiener, *Cybernetics*.

As early as 1948, founder of cybernetics Norbert Wiener used the distinction between Newtonian time, which is absolute and reversible, and Bergsonian time, which is experiential and irreversible, to posit that cybernetic *governors* operate in Bergsonian time along with living organisms.³² Analyzing Wiener's cybernetic governor through an actor's chronoception, which we developed throughout our framework, enriches the cybernetic formulation in three following ways.³³ First, following section 2, a system can change much faster or much more slowly than the governor can perceive and thus control for. Then, following section 3, computation enables the governor to expand its chronoception and potentially its control to different timescales. Finally, according to section 4, the evolution of systems as nestings of temporalities develops in a complex span of entangled timescales, whose interference with the timescale of an actor produces a misalignment that promotes temporal adaptation.

The problem of cybernetic governance, posed at the interplay of these temporal dynamics, becomes critical for our analysis. On the one hand, it makes clear that *homeostasis*—or optimal regulation of a target system—is often computationally intractable. On the other hand, it hints at the fact that homeostasis can be less beneficial, as by adapting to the interplay of temporalities an actor may transcend into another state that is much more beneficial than the one it was in before. This is what the Cybernetic Culture Research Unit (CCRU) emphasized as an antithesis to the goal of homeostasis.³⁴ What if, instead of entertaining negative feedback loops that help regulate and keep a system in its initial state, we could entertain a positive feedback loop that would help translate a system into another mode of existence that is much more beneficial? This thought developed into a positive theory toward technology known as *accelerationism*.

This turn can also be explained through the technological breakthroughs of the current century. While Wiener's theories were inspired by Black's and Maxwell's feedback systems, which later influenced Bellman's theory of optimal control, the CCRU realized the scale and intensity of digital technologies and was inspired by discoveries in chaos theory and connectionist models, focusing on the potential of nonlinear dynamics inside cultural and technological systems. However, the partial inspiration of new technological advancements tends to always inspire folk theories about the nature of the world, which are inclined to focus on the parts that were ignored in previous ontologies. In this way, the political debate between homeostasis and acceleration becomes reminiscent of the one between the politics of cooperation of Kropotkin and the politics of competition of Francis Galton, which was popular around the turn of the previous century. Yet, from a thorough study of existing real-world systems, it becomes apparent that they cannot be neatly divided into those profiting from either homeostasis or acceleration. For example, while some mechanisms of the human body are homeostatic and rely on negative feedback, like digestion, others rely on positive feedback. In a classic example, a baby pushes out of its womb after a specific time, during which it has developed enough to do so, triggering the biological process of birth. Important here is that the dynamics of one system scaffold on top of the dynamics of another. Birth codifies the temporal barrier that functions as a statistical guarantee that a baby has matured enough to survive the world after exiting the womb. In other words, the accelerated state transition of birth is what guarantees homeostasis.

Under that lens, a strict political divide between an (essentialist) politics of homeostasis and a (transcendental) politics of acceleration seems futile. Our final goal is thus to offer a change in perspective. Our framework reveals that we could focus instead on the binding element between these dynamical systems: computation. Instead of conceptualizing the ultimate cybernetic category of *teleogenesis*—the creation of a system that caters its own goals—as a problem of control, we can think of it as a problem of chronoceptual alignment. While systems both depend on and resist the exploitation of one another, they lack knowledge of the effect of their actions. Instead of placing computation at the center of cybernetic governance, we should think of computation as the procedure that allows different actors within an ecosystem to connect, to model one another. Computation becomes the chronoceptual expansion of the world onto itself.

Ultimately, our analysis reframes the problem of cybernetic governance from a problem of control to a problem of perception. Instead of the thesis of homeostatic control leading to artificialization, of nature transformed into agriculture, and its antithesis of accelerated technological expansion leading to extinction in a future outlined by climate change, what if we could propose a synthesis that sees computation as the *homeostatic acceleration* of the medium between individual governors? Computation is what brings the dynamic changes brought by the action of an actor back to its own timescale. By trying to solve them, the actor can both evolve and endure. Dissolved across and into nature, computation can assemble sensing and infrastructure technologies to help systems better learn to adapt to their environment. Instead of being treated as the material expansion of the world in the form of sensing, from soil sensors to satellites, computation can be treated as that which provides the proper ground truth driving the recursive simulations of individual actors toward an enduring evolution.

³² Wiener, *Cybernetics*.

³³ Denizhan, "Intelligence as a Border Activity."

³⁴ Cybernetic Culture Research Unit, Writings 1997–2003.

6 Conclusion

In this work, we have proposed a new framework for understanding how technology enters our relationship with the social and natural world. We have done this through the lens of time, moving beyond the conventional view that limits technology to a disruptive force, accelerating or resisting the arrow of time. Instead, we focused on how technology mediates our perception of time, which in turn alters the terms of interaction with our environment.

At the core of this framework lies a fundamental tension—one between chronoconceptual speed, the rate at which an actor perceives change, and characteristic speed, the rate at which their environment evolves. The interplay of these two timescales defines the conditions of engagement: where their mismatch is too great, the temporal filter prevents meaningful interaction; where they align, reciprocity is possible. This region is named the *zone of learnability* and is where adaptation and dynamic interaction between environment and actor is permitted. It is here that technology operates—not as an exogenous metronome, but as a system that modulates chronoception, expanding the zone of learnability.

As computation extends chronoception, it makes it possible for actors to engage with broader timescales of characteristic time. Yet this expansion is not neutral—actors do not simply perceive but actively affect the environments they interact with. In fact, both actors and environments consist of nested and interwoven temporalities of further systems, which form profiles of characteristic time. Their intersection guides temporal adaptation, where technology acts as a form of coordinating medium across temporalities.

Our framework has implications for the natural conceptualization of the framework of cybernetics, where computation is traditionally the means of governance of an actor over another system. Instead of focusing on their feedback dynamics, which we conceptualize as homeostasis and acceleration, our analysis shifts the focus to a problem of perception: technology mediates and coordinates systems whose phenomena affect complex, far-reaching timescales. As Wiener would have expected, time becomes the element that renders computation the common thread across nature and culture.

Bibliography

- Aday, Jacob S., Julia R. Wood, Emily K. Bloesch, and Christopher C. Davoli. "Psychedelic Drugs and Perception: A Narrative Review of the First Era of Research." *Reviews in the Neurosciences* 32, no. 5 (2021): 559–71. <https://doi.org/10.1515/revneuro-2020-0094>.
- Appleton, Joyce. "The Gut-Brain Axis: Influence of Microbiota on Mood and Mental Health." *Integrative Medicine (Encinitas)* 17, no. 4 (2018): 28–32.
- Arianoutsou-Faraggitaki, Maria. "Desertification by Overgrazing in Greece: The Case of Lesvos Island." *Journal of Arid Environments* 9, no. 3 (1985): 237–42. [https://doi.org/10.1016/S0140-1963\(18\)31325-9](https://doi.org/10.1016/S0140-1963(18)31325-9).
- BBC Earth. "Snakes Hunt Bats in a Cave. | Planet Earth | BBC Earth." March 29, 2017. YouTube, 2:21. <https://www.youtube.com/watch?v=UbqVF2qaAcE>.
- Bergson, Henri. *Time and Free Will: An Essay on the Immediate Data of Consciousness*. Routledge, 2014.
- Bratton, Benjamin H. *The Stack: On Software and Sovereignty*. MIT Press, 2016.
- Carey, James W., and Gregory S. Adam. *Communication as Culture, Revised Edition: Essays on Media and Society*. Routledge, 2008.
- Castells, Manuel. "Space of Flows and Timeless Time." In *Communication Power*. Oxford University Press, 2009.
- Colville, Robert. *The Great Acceleration: How the World Is Getting Faster, Faster*. Bloomsbury Publishing, 2016.
- Comandon, Jean, dir. *La croissance des végétaux*. Institut Pasteur, CERIMES, 1929. Canal-U. <https://doi.org/10.60527/sw7k-cv79>.
- Cybernetic Culture Research Unit. *Writings 1997–2003*. Urbanomic, 2015.
- Denizhan, Yagmur. "Intelligence as a Border Activity Between the Modelled and the Unmodelled." *Angelaki* 28, no. 3 (2023): 25–37. <https://doi.org/10.1080/0969725X.2023.2216542>.
- Eagleman, David M. "Brain Time." In *What's Next: Dispatches from the Future of Science*, edited by Max Brockman. Vintage Books, 2009.
- Eddington, Arthur. *The Nature of the Physical World: The Gifford Lectures 1927*. Vol. 23. BoD—Books on Demand, 2019.
- Gable, Philip A., Amanda L. Wilhelm, and Brandon D. Poole. "How Does Emotion Influence Time Perception? A Review of Evidence Linking Emotional Motivation and Time Processing." *Frontiers in Psychology* 13 (April 2022): 848154. <https://doi.org/10.3389/fpsyg.2022.848154>.
- Gilbert, Scott F. "Holobiont by Birth: Multilineage Individuals as the Concretion of Cooperative Processes." In *Arts of Living on a Damaged Planet: Ghosts and Monsters of the Anthropocene*, edited by Anna Tsing, Heather Swanson Elaine Gan, and Nils Bubandt. University of Minnesota Press, 2017.
- Healy, Kevin, Lauren McNally, Graeme D. Ruxton, Natalie Cooper, and Andrew L. Jackson. "Metabolic Rate and Body Size Are Linked with Perception of Temporal Information." *Animal Behaviour* 86, no. 4 (2013): 685–96. <https://doi.org/10.1016/j.anbehav.2013.06.018>.
- Heidegger, Martin. *Being and Time*. SUNY Press, 2010.
- Intergovernmental Panel on Climate Change. *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by Hoesung Lee and Jim Skea Romero. IPCC, 2023. https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_FullVolume.pdf.

- Kuijter, Lenneke, and Matthias Laschke. "Designing for a Post-Growth Society Through the Eco-Harmonist: A Critical Examination of the Role of HCI and Technology Design." In *NordiCHI '24: Proceedings of the 13th Nordic Conference on Human-Computer Interaction*. Association for Computing Machinery, 2024. <https://doi.org/10.1145/3679318.3685405>.
- Latour, Bruno. *Facing Gaia: Six Lectures on the Political Theology of Nature*. Gifford Lectures at the University of Edinburgh, 2013.
- Mach, Ernst, and Peter Salcher. "Photographische Fixirung der durch Projectile in der Luft eingeleiteten Vorgänge." *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften* 95 (1887): 764–80. <https://doi.org/10.1002/andp.18872681008>.
- Marx, Karl. *Capital: Volume 3: A Critique of Political Economy*. Penguin, 1981. Originally published 1894.
- McFall-Ngai, Margaret. "Noticing Microbial Worlds: The Postmodern Synthesis in Biology." In *Arts of Living on a Damaged Planet: Ghosts and Monsters of the Anthropocene*, edited by Anna Tsing, Heather Swanson Elaine Gan, and Nils Bubandt. University of Minnesota Press, 2017.
- McLuhan, Marshall. *Understanding Media: The Extensions of Man*. MIT Press, 1994.
- Muybridge, Eadweard. *The Horse in Motion*. Leland Stanford, 1878.
- Nielsen, Jakob. "Time Scales of UX: From 0.1 Seconds to 100 Years." *UX Tigers*, November 24, 2024. <https://www.uxtigers.com/post/time-scales-ux>.
- Rosa, Hartmut. *Social Acceleration: A New Theory of Modernity*. Columbia University Press, 2013.
- Staudacher, Claus. "Simmel's Sociology of Time: On Temporal Coordination and Acceleration." *Time & Society* 32, no. 2 (2023): 210–31. <https://doi.org/10.1177/0961463X231161401>.
- Tan, Xi, and Min Hu. "The 'Speed and Passion' of the Short Video Era." In *Proceedings of the 2022 3rd International Conference on Mental Health, Education and Human Development (MHEHD 2022)*. Atlantis Press, 2022. <https://doi.org/10.2991/assehr.k.220704.220>.
- Thompson, Edward P. "Time, Work-Discipline, and Industrial Capitalism." In *Class: The Anthology*, edited by Stanley Aronowitz and Michael J. Roberts. Wiley-Blackwell, 2017. <https://doi.org/10.1002/9781119395485.ch3>.
- Virilio, Paul. "The Illusions of Zero Time." In *Virilio and Visual Culture*, edited by John Armitage and Ryan Bishop. Edinburgh University Press, 2013.
- Wang, Xiaojuan, Janne Kontkanen, Brian Curless, et al. "Generative Powers of Ten." Preprint, *arXiv*, December 4, 2023. <https://doi.org/10.48550/arXiv.2312.02149>.
- Wark, McKenzie. *Capital Is Dead: Is This Something Worse?* Verso Books, 2019.
- West, Geoffrey. *Scale: The Universal Laws of Life, Growth, and Death in Organisms, Cities, and Companies*. Penguin, 2018.
- Wiener, Norbert. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT Press, 2019.
- Winsberg, Eric. "Computer Simulations in Science." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. Stanford University, 2013. <https://plato.stanford.edu/archives/sum2013/entries/simulations-science/>.
- Wortel, Meike T., Han Peters, Juan A. Bonachela, and Nils C. Stenseth. "Continual Evolution Through Coupled Fast and Slow Feedbacks." *Proceedings of the National Academy of Sciences* 117, no. 8 (2020): 4234–42. <https://doi.org/10.1073/pnas.1916345117>.
- Zilber-Rosenberg, Ilana, and Eugene Rosenberg. "Role of Microorganisms in the Evolution of Animals and Plants: The Hologenome Theory of Evolution." *FEMS Microbiology Reviews* 32, no. 5 (2008): 723–35. <https://doi.org/10.1111/j.1574-6976.2008.00123.x>.



5 Mimesis of Mimesis

The role of representation within machine cognition remains as contentious as it has historically been in discussions of evolved human cognition—perhaps even more so. Central to this debate is whether artificial intelligences genuinely possess concepts, and if they do, whether these concepts constitute authentic forms of representation.

This issue transcends theoretical curiosity, becoming critically relevant as AI systems, particularly those grounded in language—a fundamental human representational capacity—become increasingly embedded in our infrastructural reality. Cognition itself has become infrastructural, and representational thought, by extension, occupies a similarly foundational status. Yet, if human symbolic culture originates from mimetic processes, does the artificial replication of these processes represent a straightforward “mimesis of mimesis,” or does it signify something qualitatively distinct?

Exploring this question raises another layer of complexity: Perhaps creating models of our models will illuminate the underlying dynamics of representation, or perhaps such recursive approaches will only defer definitive answers into an infinitely fractal conceptual space.

Alternatively, practical experimentation with these representations of representations may yield more tangible insights. By employing artificial representations as tools in developing novel cultural practices, we may witness a narrowing rather than a widening of the gap between signifier and signified. Yet, this collapse of symbolic distance poses its own challenges: Is a closer fusion of representation and meaning inherently beneficial, or could it lead to unforeseen complications?

These projects navigate these intricate philosophical and practical landscapes, addressing how evolving machine cognition reshapes our fundamental understanding of representation, symbolism, and cultural dynamics.

5a *Minimum Viable Interiority*

Non-player characters (NPCs) in video games are often highly functional agents that interact seamlessly with their virtual environments, yet they typically lack even the illusion of an inner experiential life. This raises an intriguing philosophical question: Are NPCs philosophical zombies—entities indistinguishable from conscious beings in outward behavior but devoid of subjective experience? Philosophical zombies, or *p-zombies*, serve as conceptual tools to explore consciousness, defined precisely by their absence of interiority, or the encapsulation of cognitive states and processes unobservable and unpredictable from external viewpoints.

NPCs, thus characterized, have not achieved genuine individuation; they remain integrated parts of a broader functional manifold without self-contained inner states. To investigate the minimal criteria required for genuine interiority, this project employs a pandemonium architecture—an approach inspired by cognitive models where multiple independent subagents compete or cooperate to produce coherent behaviors. Such architectures mirror, at a simplified level, the structure of biological brains, where interiority emerges from the coordinated activity of cortical columns and neuronal networks.

Through this model, this research seeks to define and simulate the minimum viable conditions necessary for genuine interiority. A central conclusion emerges: While interiority undeniably involves multiple interacting subagents, the phenomenon itself critically requires closure—a boundary or encapsulation distinguishing internal processes from external observation. Precisely identifying the locus and nature of this closure is fundamental to understanding how genuine interiority can be realized computationally. Ultimately, this project contributes to ongoing philosophical and cognitive inquiries by clarifying the distinctions between mere functional agency and authentic interiority, thereby advancing our understanding of consciousness and individuation in both biological and artificial agents.

5b *Generative Topolinguistics*

A large language model (LLM) can be understood as a hypergraph, a mathematical and spatial formalization of the intricate semantic relationships connecting words and ideas within a language. At its core, an LLM spatializes language through embeddings—vectors assigning words specific coordinates within a complex topological geometry. These embeddings do not merely reflect linguistic structure; they encode deeper sociolinguistic dynamics, illuminating how sociality itself is geometrically constituted and maintained.

Traditionally, embedding visualizations serve primarily as static maps, passively depicting semantic relations. This project proposes a radical inversion: utilizing embedding visualizations not just as passive representations but as active interfaces capable of generating novel semantic outputs. In other words, rather than merely reflecting existing linguistic structures, embedding visualizations can become dynamic tools to shape and manipulate the semantic space itself, actively influencing sociolinguistic evolution.

This project proposes a series of experiments to systematically manipulate these semantic topologies, aiming to investigate how structured interventions in embedding spaces can yield emergent, interpretable sociolinguistic phenomena. By intentionally shaping semantic geometry, it uncovers higher-order insights into language's social fabric—how meanings propagate, evolve, and influence collective understandings and interactions.

At a deeper conceptual level, this exploration grapples with a provocative recursion: if language represents reality, and embeddings represent language, embedding visualizations become representations of representations of representations. By navigating and intervening in this layered structure, the project opens possibilities for novel forms of linguistic agency, enabling new approaches to understanding and influencing the complex interplay between language, thought, and society.



Minimum Viable Interiority

Iulia Ionescu

University of the Arts
London

Murad Khan

University of the Arts
London

Alasdair Milne

Serpentine & King's College
London

Cezar Mocan

Artist

Abstract

Debates about collectivity have become increasingly prevalent across computational and philosophical approaches to the modeling of intelligent systems. This paper explores whether these prevailing conceptions of collectivity adequately account for the “individual” as it emerges in the context of AI applications, which consist of distributed systems coordinating to give the appearance of a unified agent. Taking collective intelligence as a given, our thought experiment explores a functionalist approach to the construction of the individual, focusing on the feature of minimum viable interiority as a necessary precondition for cohering a model of collective intelligence from the bottom up. Building on functionalist experiments from p-zombies to non-player character design, we leverage Oliver Selfridge’s “pandemonium architecture” to construct a theory of functional closure suited to explain the mechanisms under which a unified individual emerges from a collective. We propose a speculative application of this theory that utilizes DeepMind’s Concordia library, schematizing an experimental framework under which interiority is established as an emergent phenomenon of functionally closed systems. Contrary to prevailing theories of collective intelligence, we argue that, rather than the collective being greater than the sum of its individuals, the individual is greater than the sum of its collectives. Such an individual, when composed of functionally closed collectives, is contradistinguished from open collectives such as flocks or swarms, often deemed synonymous with collective intelligence.

Keywords

interiority; collective intelligence; philosophy of mind; agent-based modeling; non-playing characters; multi-agent systems; functionalism

1 Introduction: From the Individual to the Collective (and Back Again)

“Collective intelligence” has become a popular explanatory paradigm across disciplines, applied to many complex phenomena. Sometimes conflated with “swarm” intelligence, its explanations include behavioral biological systems,¹ human systems,² and even technical systems.³ In particular, collective intelligence holds promise for frontier paradigms in artificial intelligence such as foundational language models. This theory accounts for the sheer scale of human agents that such models rely on, as well as gesturing toward the distributed nature of AI infrastructure that tends to mesh awkwardly with nominally anthropocentric framings of the individual. This perspective is perhaps best steelmanned by turning to Falandays and colleagues’ argument that “all intelligence is collective intelligence.”⁴ These authors challenge traditional notions of individual cognition and agency, suggesting that intelligence emerges from the interactions of distributed systems rather than residing within a singular, bounded entity. Here, we take this a step further by posing an epistemological question: After the turn toward collective intelligence, to what extent does the individual still remain a tool for providing insights into the ontological questions of agency?

We respond to this growing consensus by accepting its proposition as true, and then running a counterfactual: If intelligence is indeed collective all the way down, what would be required to engineer an individual from scratch? In other words, how would we reconstruct a functionally *singular* entity from the *multiple* components of intelligence? By establishing a counterfactual thought experiment in which we re-constitute the individual on the basis of collectivity, we explore the extent to which the concept of agency can be reworked for frontier AI systems that orchestrate multiple agents across sociotechnical domains. To this end, we do not aim to offer an explanation of intelligence, nor to discover the locus of “mind,” or tender a claim pertaining to the “hard” qualities of mind such as sentience or consciousness. Instead, we review the integrity of the individual agent, taken as an entity that acts through the specific feature of *interiority*, contradistinguished from those aforementioned qualia-bearing designations of the mind. When working within this realm of action between agents, interiority consists of a form of privileged access to internal states that drive action. In doing so, we consider the possibility that synthetic agents might develop not only to be “black boxed” to outside observers but also to preclude reflexive insight into their own internal operational logic.

Initially, we suppose that some variety of encapsulation might be a necessary (though insufficient) precondition to interiority, and that forms of privileged access to one’s interior states is, in fact, an emergent phenomenon that motivates decision-making, or otherwise “agentic” behaviors. Investigating encapsulation as a preliminary notion opens the possibility that there may be intrinsic dynamics essential to individuals that are salient to the explanation of group dynamics. Furthermore, some of these dynamics cannot be accessed by external observers through mere behavioral observation, thus requiring explanation at a different analytical level. To investigate this problem we propose a thought experiment, followed by an initial computational version using DeepMind’s Concordia library, an agentic framework primed for experiments in social interaction. Our experiment contrasts a classic schematic of inter-agent (often termed *multi-agent*) interaction against what we term *intra-agent intra-action* to denote the information transfers that occur *within* encapsulated agents.⁵ A diagrammatic armature for this experiment can be found in section 4.

2 Engineering Interiority

What thinking actualizes in its unending process is difference

— Hannah Arendt, *Life of the Mind*

Interiority is a minor concept in the philosophy of mind. Thomas Duddy argues for the efficacy of the term, despite its seemingly fatal association with Cartesian dualism.⁶ A near-consensus, from schools of thought as divergent as eliminative materialism and poststructuralism, amounts in Duddy’s view to a “bias [that] has inhibited progress towards adequately complex concept[s] of mind and self.”⁷ For Duddy, the duality of interior and exterior cannot be reduced to that of mind and body, but is in fact explanatorily necessary as part of a more holistic, “post-Cartesian” view of the mind.

One touchstone that complicates the monistic integrity of that interiority might be Hannah Arendt’s figure of the “two-in-one,” which characterizes the internal dialogue we engage in with ourselves.⁸ Arendt provides a model of interiority that is necessarily relational: “It is this duality of myself with myself that makes thinking a true activity, in which I am both the one who asks and the one who answers. Thinking can become dialectical and critical because it goes through this questioning and

¹ Beekman et al., “Biological Foundations.”

² Rosenberg, “Artificial Swarm Intelligence.”

³ Lévy, *Collective Intelligence*.

⁴ Falandays et al., “All Intelligence.”

⁵ Here we acknowledge Barad’s coinage of “intra-action” (Barad, *Meeting the Universe Halfway*, 33), but the present argument attempts to derive a parallel conception of the same term.

⁶ Duddy, *Mind, Self and Interiority*.

⁷ Duddy, *Mind, Self and Interiority*.

⁸ Arendt, *Life of the Mind*.

answering process.”⁹ This complex interiority is obscured for the purposes of inter-action: “Certainly when I appear and am seen by others, I am one; otherwise I would be unrecognizable.”¹⁰ Interiority is defined by a mechanism which is intra-active, obscured from the other for whom this appearance exists.

Building on Duddy’s critique, we reframe the concept of interiority in functionalist terms,¹¹ asking what minimal conditions must be met for an entity to possess a form of interior operation distinct from its external behaviors. This approach builds on the perspective posited by John Macmurray,¹² who argues that the individual is fundamentally constituted through action and relation rather than through introspection. Though we seek to test the interiority at first as a function of encapsulation, the permeability afforded by conceiving of the individual as an *actor* rather than merely a *thinker* lays the foundation for an actor that permeates the edge of the individual without necessitating the individual be a cognizant subject.

The notion of actors that are not thinkers has since been run to its logical extreme across both philosophy and game studies. David Chalmers’s philosophical zombie (or *p-zombie*) thought experiment, while traditionally positioned as a challenge to functionalist accounts of consciousness, offers a productive starting point for our functionalist investigation of interiority. The *p-zombie*—a being behaviorally identical to a conscious human but lacking subjective experience¹³—helps us define the theoretical minimum from which interiority might emerge. Rather than accepting the thought experiment’s anti-functionalist implications, we repurpose it to explore how increasing levels of functional complexity and organization might bridge the gap between purely mechanical behavior and a minimal form of interiority. This approach allows us to ask: What minimal architectural conditions must be added to a *p-zombie*-like system to test for the existence of interiority?

Video game non-player characters (NPCs) likewise offer a prototypical elaboration of the *p-zombie* concept within digital environments. NPCs are computer-controlled entities designed to populate virtual worlds and enhance player immersion through the simulation of realistic behaviors, including appearance, movement, dialogue, and decision-making.¹⁴ While primarily fulfilling practical roles—such as providing challenges, services, loot, or narrative direction—NPCs embody a key characteristic of *p-zombies*: They exhibit behaviors that evoke those of a conscious being while lacking the features of genuine awareness or subjective experiences that we would expect from the former. Just as *p-zombies* respond to stimuli and interact with their environment in ostensibly appropriate ways, NPCs operate through the execution of preprogrammed routines, responding to in-game events or adhering to predefined scripts. This mechanistic underpinning of NPC behavior provides a tangible, although virtual, manifestation of the *p-zombie* construct. In the context of game design and player experience, for the interacting player a well-crafted NPC should, ideally, be indistinguishable from human-controlled characters—a principle substantiated by numerous studies on NPC believability,¹⁵ mirroring the behavioral indistinguishability central to the *p-zombie* thought experiment.

Language models offer an even more sophisticated instantiation of the *p-zombie* concept than traditional NPCs. While maintaining the core characteristic of exhibiting intelligent behavior without the guarantee of any “hard” qualities of mind, large language models demonstrate unprecedented capabilities in natural language interaction, abstract reasoning, and even apparent self-reflection.¹⁶ When used to power NPCs, these models create agents that can engage in open-ended dialogue, demonstrate contextual awareness, and maintain consistent personas across interactions. This combination of sophisticated behavior with uncertain internal states makes language model-based agents particularly valuable for studying the construction and emergence of interiority.

We argue that both traditional NPCs and language model-based agents, as quasi-material instantiations of *p-zombies*, offer experimental shells from which to build and observe the emergence of interiority from the ground up. The *NPC-zombie* then becomes an experimental philosophical subject for analysis. Within the controlled environments of video game worlds, we can systematically manipulate variables and observe outcomes, establishing a simplified yet precise context for studying agent behavior. The observable and quantifiable nature of these artificial agents facilitates an empirical analysis of the relationship between internal processes and external actions. Drawing on the existing body of research in game AI, particularly the extensive work on NPC design and implementation,¹⁷ this approach is well-positioned to advance our understanding of the minimal conditions necessary for interiority. Moreover, the scalable complexity of NPC cognitive architectures allows for a gradual approach to constructing interiority, progressing from simple behavioral models to more sophisticated cognitive frameworks.

This scalability suggests a path toward understanding how collective intelligence might be encapsulated within individual agents, where the individual emerges as a container for multiple

⁹ Arendt, *Life of the Mind*, 185.

¹⁰ Arendt, *Life of the Mind*, 183.

¹¹ Pollock, *Build a Person*.

¹² Macmurray *Self as Agent*.

¹³ Chalmers, *Conscious Mind*, 94–96.

¹⁴ Lankoski, “Character Design Fundamentals.”

¹⁵ Warpefelt and Verhagen, “Non-Player Character Believability.”

¹⁶ Bommasani *et al.*, “Opportunities and Risks”2022; Piché *et al.*, “LLMs Can Learn Self-Restraint”2024; Renze and Guven, “Self-Reflection in LLM Agents.”2024

¹⁷ Yannakakis and Togelius, *Artificial Intelligence and Games*.

interacting processes. To investigate this emergence of individual interiority from the standpoint of the collective, we turn to cognitive architectures—particularly *pandemonium architecture*—as methodological frameworks for studying the development of bounded, yet internally complex, agents.

3 Building NPC-Zombies with Pandemonium Architecture

The evolution of NPCs in video games mirrors broader AI research trajectories. From predictable rule-based systems to more dynamic approaches like finite state machines and behavior trees,¹⁸ NPC design has increasingly focused on creating believable agents. *The Sims* popularized utility-based decision-making where characters maximize happiness by selecting actions based on personality-linked needs.¹⁹ While primarily reactive, these systems create an appearance of purposeful behavior. More sophisticated approaches like goal-oriented action planning²⁰ and cognitive architectures such as ACT-R²¹ and SOAR²² have introduced multi-step planning and modular systems for perception, learning, and reasoning. These frameworks, when adapted for NPCs,²³ produce more sophisticated agents through the integration of multiple concurrent processes vying for priority within a single decision-making entity.

The cognitive architectures described are grounded in broader cognitive science and philosophical research. Marvin Minsky posits that intelligence emerges from the interaction of numerous simple processes or agents.²⁴ Jeff Hawkins's theory proposes that the neocortex contains many distributed models of the world, each built from sensory inputs and making predictions, rather than a single hierarchical model, with these multiple models working together to form our perception and understanding of reality.²⁵ A common thread running through these approaches is the theme of multiple, parallel processes within a single agent, instantiated as needs competing for attention in the realm of action planning, possible actions competing for resources under the constraint that an agent can pursue a single action at a time, and so on. Decision-making—that determines which need to attempt fulfilling at any moment and what action plan will most likely lead to the satisfaction of that need—is a prerequisite to both interiority and intelligence. Our claim, that interiority arises as an emergent property of stacking layers of internal decision-making that the agent is not directly exposed to, is detailed in section 4 of this paper.

The idea of a tiered decision system operating on independent modules finds its most explicit expression in the “pandemonium architecture,” originally proposed by Oliver Selfridge in 1959 as a model of pattern recognition in human visual perception.²⁶ Selfridge introduced a hierarchical structure of *daemons* as simple processing units that work in parallel to analyze input data. The model consisted of multiple layers, including feature daemons that detect basic patterns, cognitive daemons that combine these features, and decision daemons that make final classifications (Figure 1).

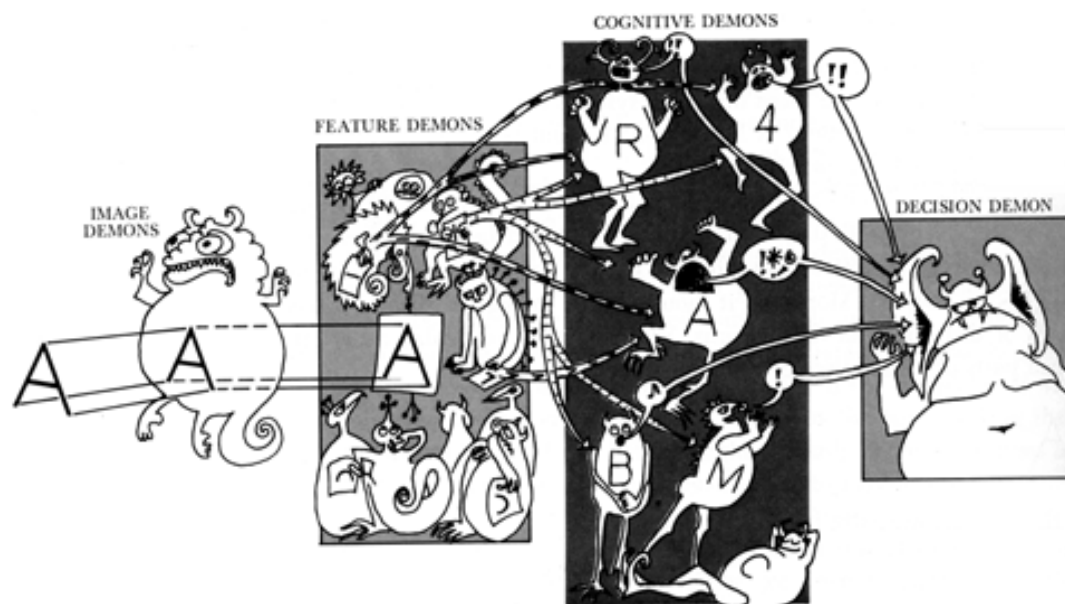


Figure 1 An illustration of Oliver Selfridge's 1959 pandemonium architecture model, drawn by Leanne Hinton. Source: Lindsay and Norman, *Human Information Processing*.

¹⁸ Buede *et al.*, “Filling the Need.”

¹⁹ Tirrell, “Dumb People, Smart Objects”; Brown, “AI Behind *The Sims*.”

²⁰ Orkin, “Goal-Oriented Action Planning.”

²¹ Ritter *et al.*, “ACT-R.”

²² Laird, *The Soar Cognitive Architecture*.

²³ Lent *et al.*, “Intelligent Agents.”

²⁴ Minsky, *Society of Mind*.

²⁵ Hawkins, *A Thousand Brains*.

²⁶ Selfridge, “Pandemonium.”

Our usage of the pandemonium architecture model in this paper extends the context of Selfridge's theory toward NPC cognitive architectures. It differs from an intuitive view of an agent as a single encapsulated model (Figure 2).

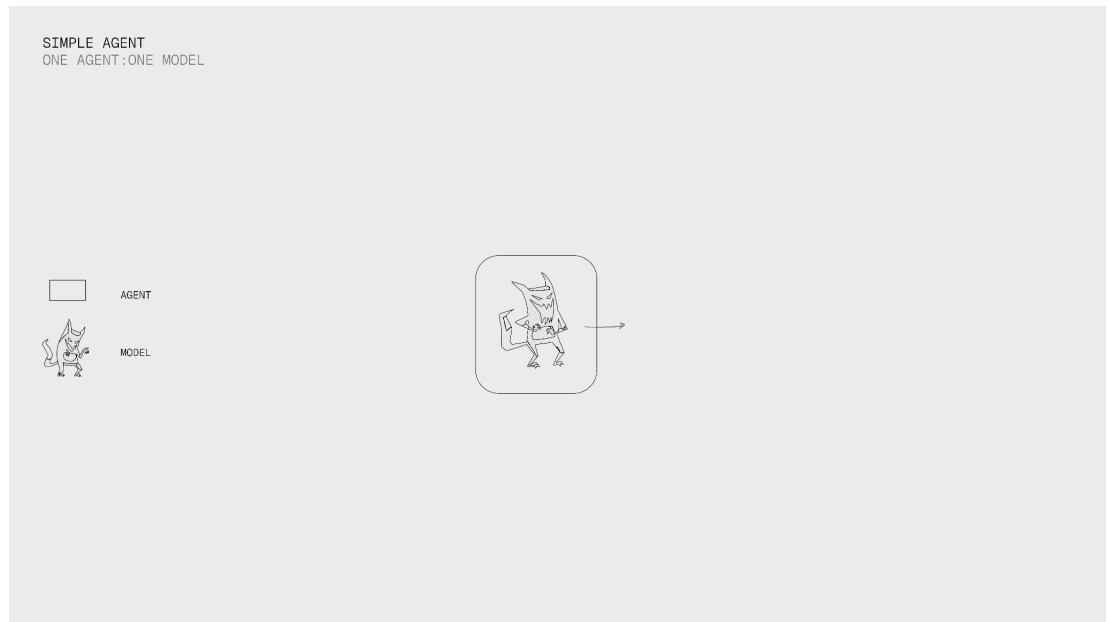


Figure 2 A simple agent composed of one model.

In contrast, in pandemonium architecture, multiple models, or *daemons*, coexist within a single agent, each processing information or generating responses independently (Figure 3).

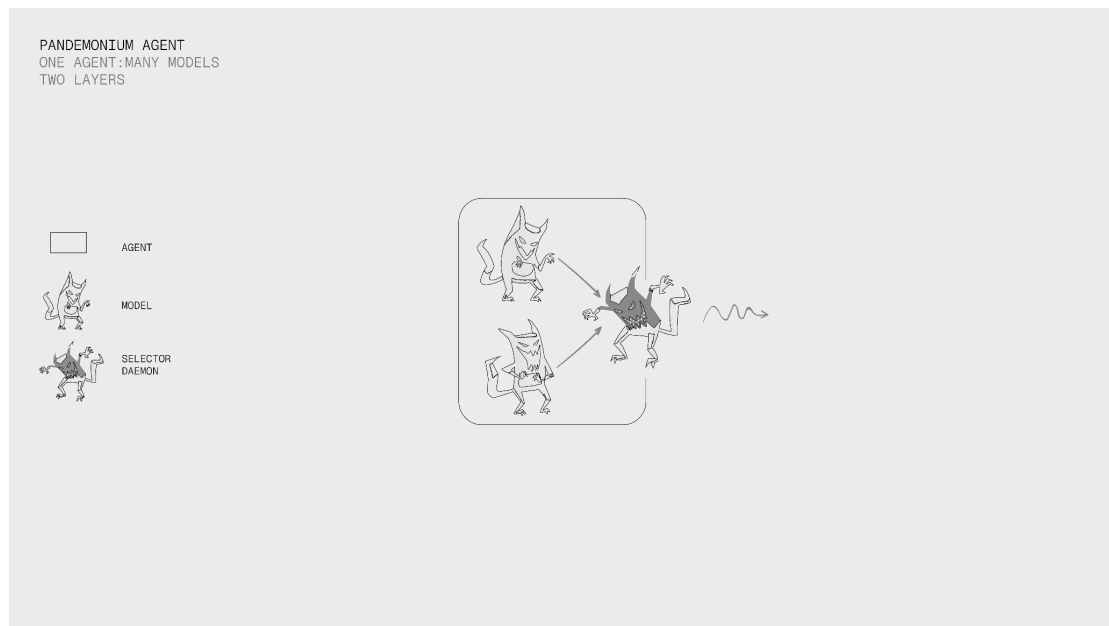


Figure 3 A pandemonium agent composed of two models and a selector daemon.

A crucial component of this architecture is the *selector daemon*, which reconciles the outputs of these competing models to generate a final action or response. This internal structure might create complexity and potentially generate more nuanced behavior, and allows us to build toward minimum viable interiority through the stacking of functionally closed layers of daemons.

4 Intra-Agent Intra-Action and Functional Closure

Although the architecture outlined by Selfridge produces an individual that equates to a single instance of pandemonium, our hypothesis focuses on how the development of interiority can be considered an emergent property of a system of nested, functional closures. We build our experimental framework on the history of organizational closure in theoretical and systems biology,²⁷ extending Alvaro Moreno and

²⁷ Maturana and Varela, *Autopoiesis and Cognition*; Moreno and Mossio, *Biological Autonomy*.

Matteo Mossio's description of "an organization of constraints" to outline a theory of *functional* closure in agent-based systems. Mossio and Moreno's framework provides an explanation of how complex biological and cognitive systems develop internal dynamics due to local constraints,²⁸ where the organization of constraints as a *collective* constitutes a system of self-maintenance. By applying this concept to agent-based systems, we can better model how artificial agents might develop collective forms of self-organizing behaviors that emerge from internal constraints rather than being solely determined by external factors. Much as Mossio and Moreno seek to expand closure from physical to biological self-maintenance, we move a step further to transpose closure to a regime of psychological self-maintenance suitable for explaining the development of interiority as a system of enclosed, privileged states.

Central to our analysis is the proposition that a functionally closed system is irreducible to a genealogical tracing of causes at each scale of operation. Rather, we hold that the distinctive feature of such a system is that each closure is causally explainable only by the events observed within its respective domain, and thus provides explanations for phenomena local to each layer.²⁹

We distinguish between two levels of interaction under these conditions: (1) *intra-agent*, defined as the dynamic interplay between *models* on a single layer of closure; (2) *inter-agent*, the engagement between functionally closed systems and across scales. Such a system exhibits closure at local scales, such that each level of operation is organized by prior bounded levels of pandemonium. A coarse overview of one such subsystem is shown in Figure 4

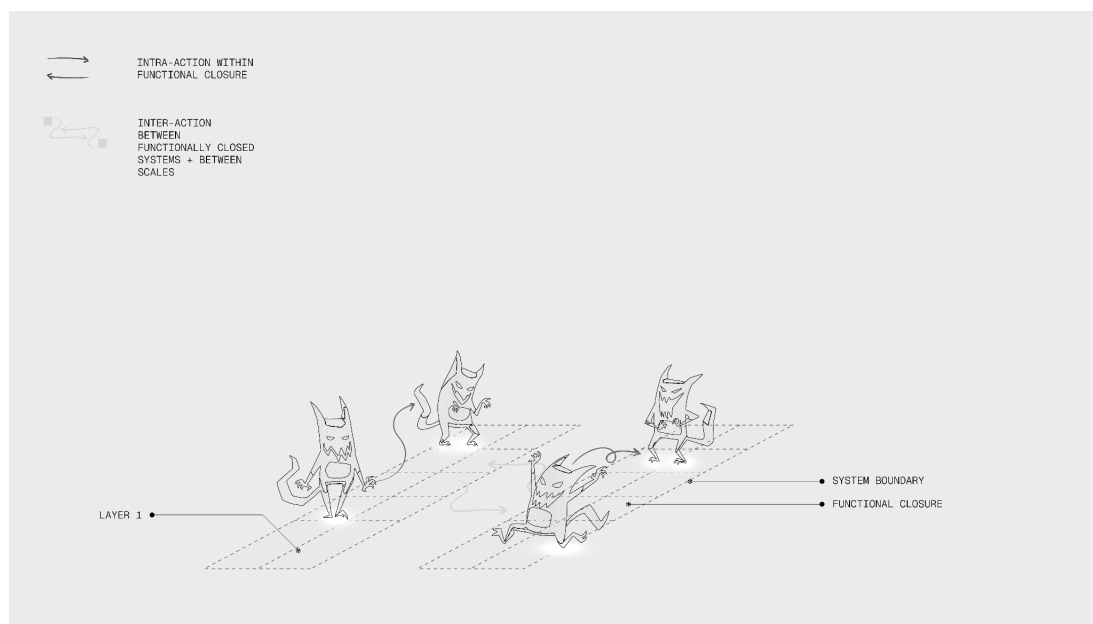


Figure 4 Inter-action and intra-action for two distinct functionally closed layers within an agent.

At this pandemonium layer, components within each functionally closed system (represented in Figure 4 as separate planes) intra-act with one another whereas closed systems inter-act with one another. Models in the lower layer are constitutive of the systems in the layer above, which emerge from the levels below (Figure 5).

²⁸ Mossio and Moreno, "Organisational Closure."

²⁹ In other words, closure is a *constitutive*, rather than *etiological*, explanation to the extent that it provides an ontic account of the development of emergent phenomena from the causal regime of constraints within a system (Salmon, *Causality and Explanation*).

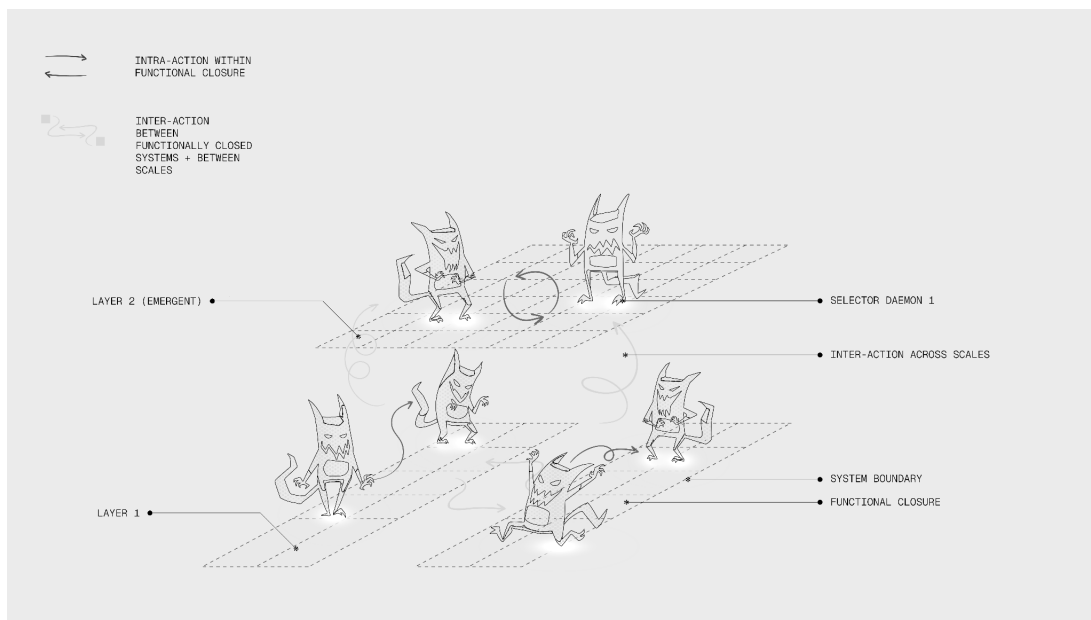


Figure 5 Inter-action and intra-action represented on a second, emergent layer.

At this pandemonium layer, components within each functionally closed system (represented in Figure 4 as separate planes) intra-act with one another whereas closed systems inter-act with one another. Models in the lower layer are constitutive of the systems in the layer above, which emerge from the levels below (Figure 5).

At this pandemonium layer, intra-action at the layer below (Layer 1) produces an emergent set of components at a higher degree of complexity. Every subsequent layer produced after Layer 1 is irreducible to the layer before. The final layer of the system encapsulates all previous layers and components and is presented as a whole (Figure 6).

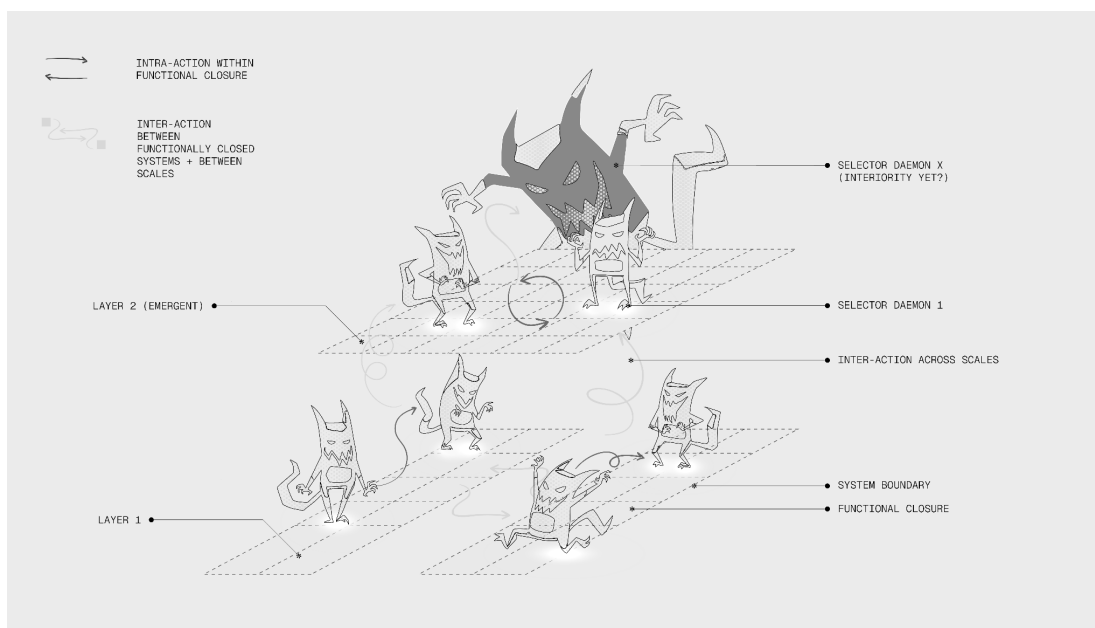


Figure 6 Top-level selector daemon (final layer) acting on the previous functionally closed layer.

4.1 Selection

Under this conceptualization of a pandemonium architecture, the role of our selector daemon is to arbitrate between intra-agent dynamics that emerge at each level of closure. To this extent, selection provides the conditions under which interiority develops as well as the process by which we come to present a state of phenomenal unity.³⁰ It is through selection that we *re-present*, or externalize, the dissonance of internal cognitive states in action in a mode perceived to be indicative of an “individual.”

³⁰ Metzinger, *Being No One*.

Building on the neuropsychological explanation provided by Michael S. Gazzaniga and Joseph E. LeDoux, which casts the left hemisphere of the brain as an executive *interpreter* of information that unifies conscious and unconscious experience,³¹ we propose that selection *interiorizes* the individual, producing a unity that is functionally taken to be the individual through the development of an emergent self-model.

4.2 Minimum Viable Interiority Constant

Interiority is therefore defined as an emergent property, partially observed by the exhibition of behavioral (ir)regularities at the level of the individual, but ultimately hidden by the bounded nature of each level of pandemonium and their eventual closure through the process of selection. Under this framework, we hypothesize that a system requires λ levels of functional closure to achieve interiority, where λ represents our *minimum viable interiority constant*. With each additional level of nesting, we add a layer of complexity, creating a decision-making hierarchy that becomes increasingly opaque to external observation and internal introspection.

Against prevailing theories of collective intelligence, which suggest that the collective is greater than the sum of its parts (individuals), we suggest the opposite: that this architecture recognizes the extent to which *the individual is greater than the sum of its collectives*. Functional closure grounds an augmented schema for Selfridge’s pandemonium architecture that necessitates the interaction of nested constraints within a single agent, where the collective self-maintenance of causal boundaries between different levels of agent interaction contributes to the global structure of interiority as an emergent property of the system.

4.3 Pandemonium Architecture Versus Neural Networks

In our proposed framework, we explore Selfridge’s pandemonium architecture as a potential model for the development of interiority. Of particular interest is the extent to which the hierarchical nature of decision-making exhibited by pandemonium agents can be refined through a theory of functional closure. Whilst it is the case that contemporary deep learning architectures, such as transformers,³² also exhibit hierarchical forms of information processing, the explicit design of interacting daemons *in pandemonium*—in which agents possess designed internal dynamics with multiple interacting component roles (where different daemons interact to produce behavior)—provides a more interpretable framework for studying the emergence of cognitive-like processes. Both approaches have their strengths and limitations in modeling cognitive processes, and future work may benefit from integrating insights from both paradigms, but these explicitly defined functional units offer a different perspective on cognitive modeling that are generative for more exploratory, conceptual research into interiority as an emergent property of a functionally closed system.

5 Experiment Design

To take this a step further, we schematize a conceptual, computational framework for testing the hypothesis that interiority—defined as a form of privileged access to internal states—can emerge in NPC agents through a pandemonium architecture. We focus on the specific dynamics of multi-model agents, each controlled by multiple internal models (or daemons) whose outputs are reconciled by layers of selector daemons. Our goal is to explore whether increasing levels of functional closure in these agents can generate what we term *minimum viable interiority*, which is characterized by complex forms of internal decision-making opaque to external observers.

We propose a software experiment using an agentic pandemonium architecture, where multiple agents operate within a simulated sandbox environment. To streamline development, we suggest building this architecture on an existing agentic framework, such as DeepMind’s Concordia. Defined as “a library to facilitate the construction and use of generative agent-based models to simulate interactions of agents in grounded physical, social, or digital spaces,”³³ Concordia is an open-source project that enables the creation of social agents driven by large language models.

The primary goal of our experiment is to observe and quantify behavioral differences between agents with varying levels of functional closure in their cognitive architectures. By creating a series of λ simulated scenarios—each identical except for the number of functional closure levels within the agents’ cognitive structures—we aim to explore how increasing levels of functional closure might contribute to the emergence of interiority, establishing a foundation for more detailed quantitative analysis in future work.

In the baseline scenario—a simple inter-agent inter-action—we simulate N agents, each with a cognitive architecture that contains a single level of functional closure: one daemon (a decision-making unit) powered by a language model, implemented as a Concordia agent. In this setup, the agents engage in inter-agent interaction but lack any intra-agent complexity. Each agent contains one internal daemon, resulting in a total of N daemons across the simulation (Figure 7).

³¹ Gazzaniga and LeDoux, *The Integrated Mind*.

³² Vaswani *et al.*, “Attention Is All.”

³³ Vezhnevets *et al.*, “Generative Agent-Based Modeling.”

In comparison, for the pandemonium configuration on the right in Figure 7 we add a second level of functional closure to each agent, implementing a basic pandemonium architecture. This consists of two first-layer daemons processing sensory input and one *selector daemon* that chooses the most appropriate response. Each agent has three internal daemons (2^2-1), resulting in a total of $3 \times N$ daemons across the simulation. In this architecture, agents begin to exhibit intra-agent intra-action, where multiple internal decision processes occur without all states being visible at higher levels or to external observers.

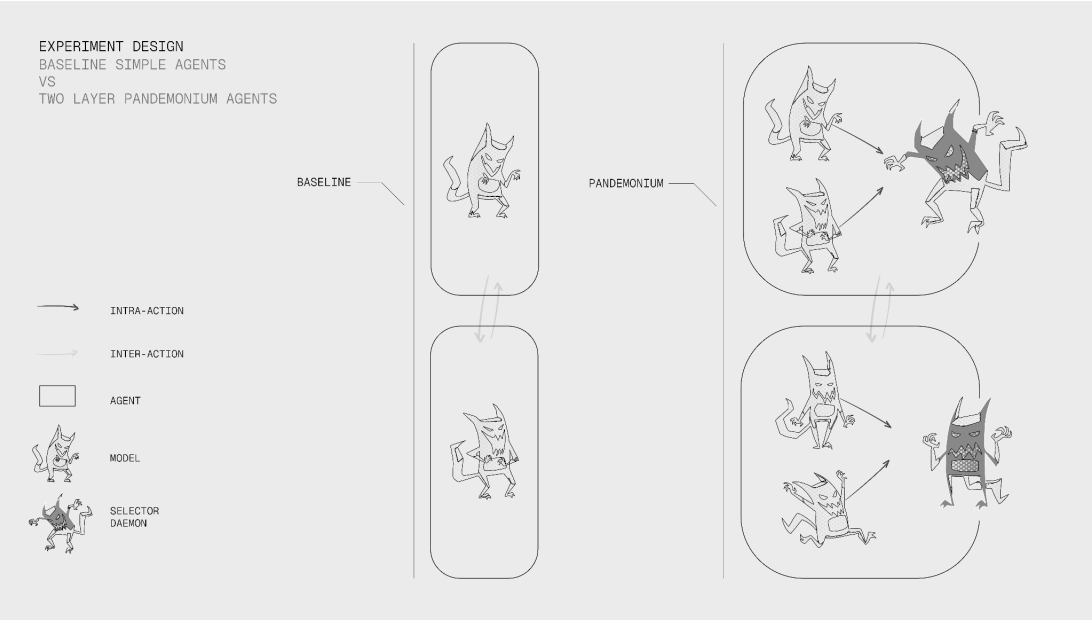


Figure 7 Comparison of the baseline scenario with a pandemonium scenario with one layer of functional closure, for $N=2$.

At a depth of three (as in Figure 8) the structure becomes more complex. The selector daemon now manages multiple second-layer daemons, each overseeing a pair of first-layer daemons. This hierarchy follows the pandemonium model, where each layer specializes in progressively abstract functions. For example, first-layer daemons might detect basic patterns in input data, while second-layer daemons integrate these patterns into more sophisticated perceptions or decisions. With three levels, each agent would have seven internal daemons, 2^3-1 , totaling $7 \times N$ daemons across the simulation.

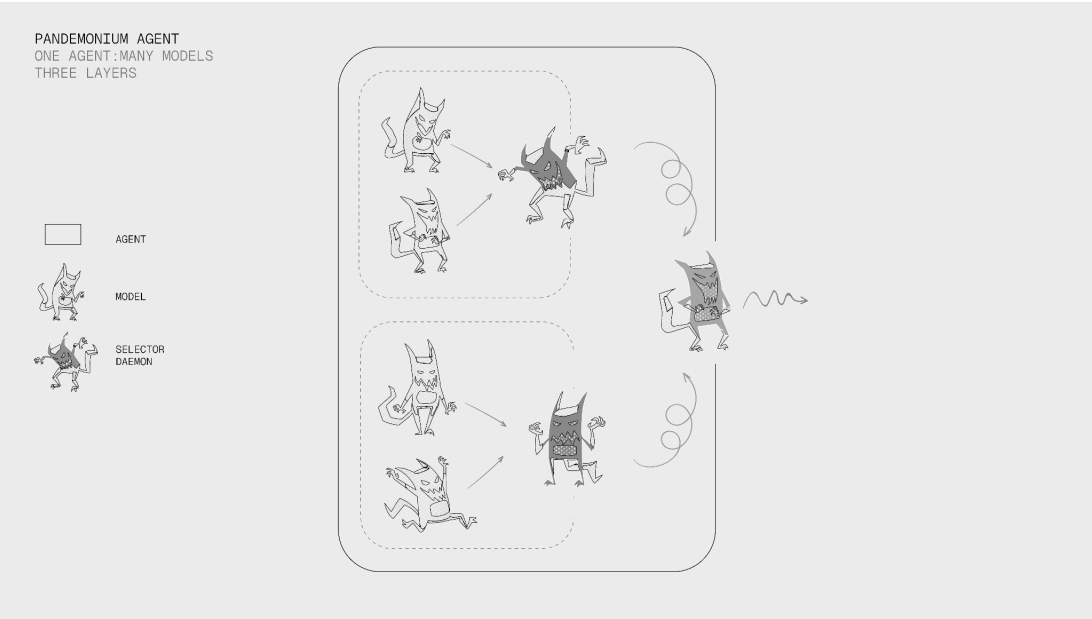


Figure 8 A single agent containing a pandemonium architecture with three layers.

As we increase the number of functional closure levels to λ , the system grows exponentially in complexity, making the higher levels more theoretical in their feasibility due to the rapidly increasing number of required daemons, $N \times (2^\lambda - 1)$ (Figure 9).

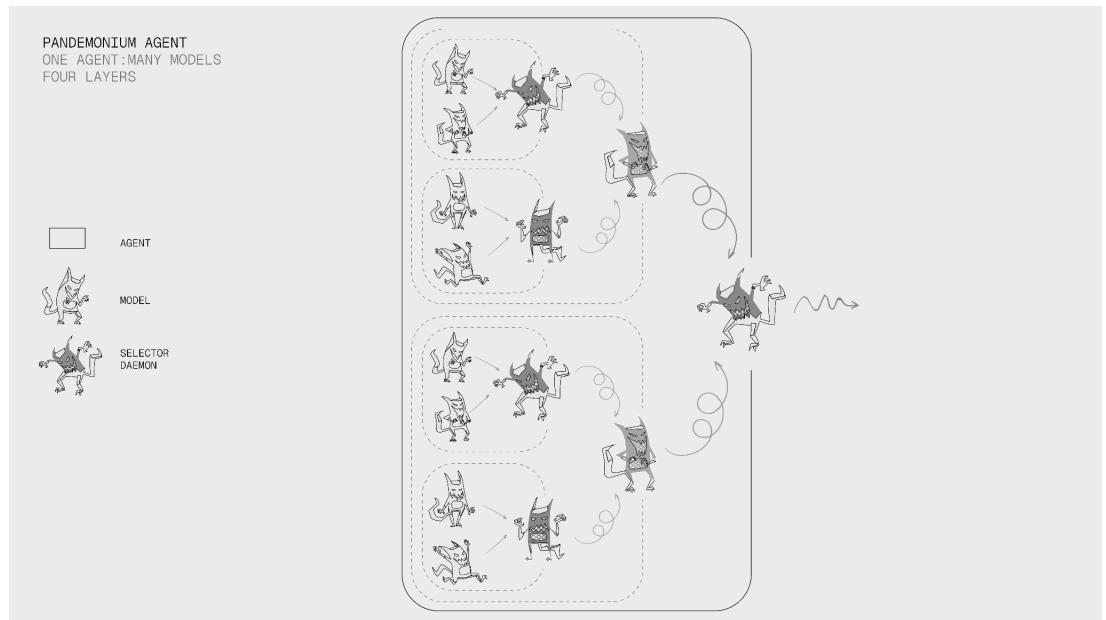


Figure 9 A single agent containing a pandemonium architecture with four layers

6 Conclusions

We have proposed a thought experiment: What could be learned by an attempt to engineer the individual from the ground up? By hypothesizing the nature of an individual as fundamentally collective, we are led by necessity to understand the organizational complexity from which an individual can emerge from a bundle of collectives. From this, we schematized a possible experiment that accounts for the emergence of a unified individual through the mechanism of functional closure. On these grounds, we pose a framework for understanding the scalar nature of interiority, a phenomenon constrained by the minimum viable interiority constant that acts as a limit to the possible regress of necessary layers.

This is not a reactionary stance against the growing consensus of collective intelligence, but rather a constructive provocation: If intelligence is indeed “collective all the way down,”³⁴ we require an adequate explanatory framework for understanding its construction across multiple scenarios. We come to the preliminary position, then, that while the individual may be composed of the *many* all the way down, it still provides an important explanatory function for collectives in which intelligence is not distributed between individual group members. In particular, if interiority emerges at λ levels of functional closure, and is displayed behaviorally, such a presentation may be anticipated as distinct from the behaviors of the swarm, where a collective has no consolidated internal functioning. This suggests a distinction between collectively-driven action where the system is functionally closed, as in an individual, and where the system is open, as in a flock or a swarm. This might lead to the reconsideration of certain collective, functionally closed systems or organizations as individuals themselves.

Our claim that interiority emerges from stacked, cascading collectives is not necessarily a critique of analyses that promote the collective as ontologically foundational. Rather, the present proposal seeks to respond to concerns around the integrity of the individual by posing a compatibilist view. The collective composes the individual, but the individual is not troubled or undermined, simply in need of reconsideration. To conclude, we reiterate our primary proposition: that against prevailing theories of collective intelligence, which suggest that the collective is greater than the sum of its parts (individuals), we suggest the opposite; that this architecture recognizes the extent to which *the individual is greater than the sum of its collectives*.

The implications of this view are better teased out through the development of empirical metrics to quantify emergent interiority in functionally closed systems, including measures of decision opacity (how predictable an agent’s behavior is from external inputs), intra-agent interaction density (the complexity of interactions between internal subsystems), and self-model coherence (the consistency of an agent’s self-representation). Such metrics could help establish when the minimum viable interiority constant (λ) is reached, potentially bridging conceptual theories of interiority with observable properties of complex AI systems.

³⁴ Falandays et al., “All Intelligence,” 1.

Further to running the proposed experiment pertaining to interiority as a minimal conception of the individual, a subsequent step might be situating the hard problem within this schema. If, as Thomas Metzinger writes, the “phenomenal self is not a thing, but a process,”³⁵ then speculation might suggest an emergent relationship between interiority and “hard” conceptions of mind at higher orders of complexity, consistent with Daniel Dennett’s view of consciousness as an emergent property.³⁶ Such emergence might be observable through various manifestations of self-modeling and self-reference in agent behavior. Regardless of one’s position on the hard problem, interiority proposes an intermediate, incremental step between the *p-zombie* and the person.

Acknowledgments

We thank Sasha Vezhnevets and Joel Leibo from DeepMind, Vinca Kruk and Daniel van der Velden from Metahaven, and fellow studio researchers, affiliate researchers and team from Antikythera. We also thank the Cosmos Institute.

³⁵ Metzinger, *Being No One*.

³⁶ Dennett, *Consciousness Explained*.

Bibliography

- Arendt, Hannah. *The Life of the Mind: The Groundbreaking Investigation on How We Think*. HMH, 1981.
- Barad, Karen. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, 2007.
- Beekman, Madeleine, Gregory A. Sword, and Stephen J. Simpson. "Biological Foundations of Swarm Intelligence." In *Swarm Intelligence: Introduction and Applications*, edited by Christian Blum and Daniel Merkle. Springer, 2008. https://doi.org/10.1007/978-3-540-74089-6_1.
- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. "On the Opportunities and Risks of Foundation Models." Preprint, *arXiv*, August 16, 2021. <https://doi.org/10.48550/arXiv.2108.07258>.
- Brown, Mark. "The Genius AI Behind The Sims." *Game Maker's Toolkit on Substack*, June 30, 2023. <https://gmtk.substack.com/p/the-genius-ai-behind-the-sims>.
- Buede, Dennis M., Bradley DeBlois, Doug Maxwell, and Beverly McCarter. "Filling the Need for Intelligent, Adaptive Non-Player Characters." In *Interservice/Industry Training, Simulation, and Education Conference*, 2013. https://www.researchgate.net/publication/292984684_Filling_the_Need_for_Intelligent_Adaptive_Non-Player_Characters.
- Chalmers, David John. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1997.
- Dennett, Daniel. *Consciousness Explained*. Penguin, 1993.
- Duddy, Thomas. *Mind, Self and Interiority*. Routledge, 1995.
- Falandays, J. Benjamin, Roope O. Kaaronen, Cody Moser, et al. "All Intelligence Is Collective Intelligence." *Journal of Multiscale Neuroscience* 2, no. 1 (2023): 169–91. <https://doi.org/10.56280/1564736810>.
- Gazzaniga, Michael S., and Joseph E. LeDoux. *The Integrated Mind*. Springer US, 1978. <https://doi.org/10.1007/978-1-4899-2206-9>.
- Hawkins, Jeff. *A Thousand Brains: A New Theory of Intelligence*. Basic Books, 2021.
- Laird, John E. *The Soar Cognitive Architecture*. MIT Press, 2019.
- Lankoski, Petri. "Character Design Fundamentals for Role-Playing Games." In *Beyond Role and Play: Tools, Toys and Theory for Harnessing the Imagination*, 139–48, 2004. <https://researchportal.tuni.fi/en/publications/character-design-fundamentals-for-role-playing-games>.
- Lent, Michael van, John Laird, Josh Buckman, et al. "Intelligent Agents in Computer Games." In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference (AAAI '99/IAAI '99)*. AAAI, 1999.
- Lévy, Pierre. *Collective Intelligence: Mankind's Emerging World in Cyberspace*. Translated by Robert Bononno. Perseus Books, 1997.
- Lindsay, Peter H., and Donald A. Norman. *Human Information Processing: An Introduction to Psychology*. Academic Press, 1972.
- Macmurray, John. *The Form of the Personal: The Self as Agent*. 2nd ed. Faber & Faber, 1966.
- Maturana, Humberto R., and Francisco J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Reidel, 1972.
- Metzinger, Thomas. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2004.
- Minsky, Marvin. *Society of Mind*. Simon and Schuster, 1988.

- Moreno, Alvaro, and Matteo Mossio. *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer, 2015.
- Mossio, Matteo, and Alvaro Moreno. "Organisational Closure in Biological Organisms." *History and Philosophy of the Life Sciences* 32, nos. 2–3 (2010): 269–88.
- Orkin, Jeff. "Applying Goal-Oriented Action Planning to Games." In *AI Game Programming Wisdom 2*, 217–28 (2003).
- Piché, Alexandre, Aristides Milios, Dzmitry Bahdanau, and Chris Pal. "LLMs Can Learn Self-Restraint through Iterative Self-Reflection." Preprint, *arXiv*, May 15, 2024. <https://doi.org/10.48550/arXiv.2405.13022>.
- Pollock, John L. *How to Build a Person: A Prolegomenon*. MIT Press, 1989.
- Renze, Max, and Elif Guven. "Self-Reflection in LLM Agents: Effects on Problem-Solving Performance." *arXiv*, May 5, 2024. <https://doi.org/10.48550/arXiv.2405.06682>.
- Ritter, Frank E., Farnaz Tehranchi, and James D. Oury. "ACT-R: A Cognitive Architecture for Modeling Cognition." *WIREs Cognitive Science* 10, no. 3 (2019): e1488. <https://doi.org/10.1002/wcs.1488>.
- Rosenberg, Louis. "Artificial Swarm Intelligence, a Human-in-the-Loop Approach to A.I." In *Proceedings of the AAAI Conference on Artificial Intelligence* 30, no. 1 (2016). <https://doi.org/10.1609/aaai.v30i1.9833>.
- Salmon, Wesley C. *Causality and Explanation*. Oxford University Press, 1998.
- Selfridge, Oliver G. "Pandemonium: A Paradigm for Learning." In *Neurocomputing: Foundations of Research*, 115–122. MIT Press, 1958.
- Tirrell, Jeremy W. "Dumb People, Smart Objects: The Sims and the Distributed Self." Paper presented at the 6th International Conference on the Philosophy of Computer Games, January 29–31, 2012, Madrid, Spain. <https://www.semanticscholar.org/paper/Dumb-People-%2C-Smart-Objects-%3A-The-Sims-and-the-Self-Tirrell/f7554fd956409cd9ba08b0dae3249d8e7a9a58cb>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. "Attention Is All You Need." Preprint, *arXiv*, last modified August 2, 2023. <https://doi.org/10.48550/arXiv.1706.03762>.
- Vezhnevets, Alexander S., John P. Agapiou, Avia Aharon, et al. "Generative Agent-Based Modeling with Actions Grounded in Physical, Social, or Digital Space Using Concordia." Preprint, *arXiv*, December 6, 2023. <https://doi.org/10.48550/arXiv.2312.03664>.
- Warpefeldt, Henrik, and Harko Verhagen. "A Model of Non-Player Character Believability." *Journal of Gaming & Virtual Worlds* 9 (2017): 39–53. https://doi.org/10.1386/jgvw.9.1.39_1.
- Yannakakis, Georgios N., and Julian Togelius. *Artificial Intelligence and Games*. Springer, 2018.



Generative Topolinguistics

Bidirectional Interfaces for Emergent Language Topologies

Iulia Ionescu
University of the Arts
London

Jenn Leung
University of the Arts
London

Yannis Siglidis
Ecole Des Ponts
ParisTech

Abstract

The experimental framework set out for generative topolinguistics seeks to investigate the sociality of meaning construction in artificial cognitive systems. While the semanticity of artificial linguistic systems is an emerging area of research, our work explores how the tokenization of language could produce new interfaces for the exploration of sociolinguistic phenomena. Generative topolinguistics presents a perspective on artificial sociality in simulated environments, employing a functionalist framework to capture its structure through token interactions inside the high-dimensional vector spaces of modern LLMs. In our model, language functions geometrically while sociality functions topologically, with changes in the topology of movement in semantic space interpreted as social behavior. Through the proposal of a bidirectional interface for large language models, we speculate how structural manipulations of semantic space could lead to the emergence of various sociolinguistic features that scaffold toward interpretable higher-order social phenomena.

Keywords

LLMs; topology; tokenization; bot-only social networks

1 Introduction

At the turn of the twentieth century, key thinkers of linguistics such as Saussure and Wittgenstein used the modeling ontologies of their time to speculate on what language could be. Their vocabularies and concepts originated from the two unbridged worlds of the classical humanities and natural sciences. Saussure's idea of language saw it as a structure, distinguishing it from *parole*, its oral manifestation, by conceptualizing it as an underlying system.¹ This view was largely aligned with developments in neuroscience at the time, where Broca's and Wernicke's areas of the brain were already found to be responsible for producing and understanding language.² However a formal connection between the two remained ambiguous. Wittgenstein focused instead on a more social aspect of language, what he called *language games*, where the meaning of a word can change through its use and interaction.³ Later, Lyotard used this concept to discuss how ideology and narrative make language almost a code that speaks for itself, encoding and recoding meaning inside the social world.⁴

Perhaps working at different scales, such theoretical approaches can be viewed in retrospect more as modeling attempts to describe discrete aspects of language. Although implementing these models to reproduce linguistic phenomena could potentially validate underlying assumptions about the nature of language, testing these methods would still require a complete framework. Instead of seeking this common underlying framework, essentialist debates between analytical models—such as the innate generative grammars of Chomsky or more experimental approaches, like the behaviorist, functionalist models of Skinner⁵—delayed the process of research due to an almost *ideological* confrontation.⁶

Computational linguistics and language modeling were efforts of the linguistic community to make such emergent ideas tangible through computation. Generative grammars were mapped to state machines,⁷ and behaviorist approaches were mapped to statistical models⁸, often n-grams.⁹ Despite their ability to create simple applications such as autocomplete, such modeling attempts were futile epistemic efforts at mapping an unreasonable or highly complex system to an analytical statistical model.¹⁰ Inspired by early models of biological neural networks, the connectionist approach grew from parts of the statistical modeling community to become the predominant modeling approach for language modeling. It converged in modeling the large statistical distribution of the sequences of subword parts, known as *tokens*, which constitute language by fitting a probability model $p(x_t | x_{\tau < t})$ to predict the next token of any sequence of a text, drawing from its history.

Developments in optimization, architecture design, and data curation enabled scaling these models to the order of trillions of network parameters, and learning from text data sets to the order of a dozen trillion (subword) tokens. Inside their weights, language got abstracted into complex and superimposed multiscale representations, some of which even learned to perform abstract algorithmic operations.¹¹ In this sense, large language models became an emergent unified model that made it possible to converge to different philosophical ideas about word interaction, structure, or self-reproducing linguistic systems through the merely empirical reproduction of written language. In other words, LLMs can challenge linguistic theories by becoming a “living proof” of what language could be. What if the philosophies of Saussure or Lyotard are now coded in some form or another in the model's parameter space, and one can now instead study them through interaction to understand their limits?

1.1 Generative Topolinguistics

When it comes to analyzing language, one of the most compelling properties of LLMs is that they map linguistic symbols into tokens, discrete chunks of vector representations, which interact and are transformed through common vector operations by learning network weight to perform next-token prediction. What makes them compelling is that inside the abstract, high-dimensional spaces occupied by such vectors, one can locate (1) structures of interactions between tokens, (2) geometric properties where vector similarity is encoded as semantic similarity, and (3) topological properties where the global structure of such token interaction can reveal patterns, which in the context of user interaction can encode sociality. There are two well-studied paradigms for studying LLMs: (1) through a top-down approach, known as representational analysis, which investigates high-level properties of the embedding

¹ Saussure, “Course in General Linguistics.”

² Ruten, “Broca-Wernicke Theories.”

³ Wittgenstein, *Philosophical Investigations*.

⁴ Lyotard, *The Postmodern Condition*.

⁵ Chomsky, *Theory of Syntax*; Skinner *Science and Human Behavior*.

⁶ Chomsky, “Case Against B.F. Skinner.” Such debates are in retrospect reminiscent of the debates in physics around the wave or particle nature of light.

⁷ Hunter, “Chomsky Hierarchy.”

⁸ Saffran, “What Is Statistical Learning.”

⁹ Shannon, “The redundancy of English.” A more complete introduction to the history of NLP can be found in Manning and Schütze, “Foundations of statistical natural language processing.”

¹⁰ Statisticians even ideologized the parameter count of their models, as in the case of Von Neumann's elephant: “With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.” See Dyson, “A meeting with Enrico Fermi.”

¹¹ Elhage et al., “Mathematical Framework.”

space; or (2) through mechanistic interventions, which identify learned algorithms of neuron–tokens interactions.¹²

Generative topolinguistics borrows from both methodologies with the purpose of designing a generative framework toward language that enables humans to understand sociolinguistic phenomena. Prior work suggests that we can not only observe but also manipulate such representations. This implies that instead of trying to model human language as a distributed embedded moving target, we can instead pose the question: Given a certain physical structure assumed by *language*, what happens to it if we interact with it by manipulating its *geometric representation*? How would that develop, *topologically*, into further interactions between artificial linguistic systems, that is, in terms of their sociality? Can new forms of sociality emerge from existing linguistic structures, and would a different language, or set of semantic relations, emerge to support new forms of sociality?

To generate answers to all these questions, we motivate and propose a bidirectional framework for analyzing and interacting with language in tokenizable space. Our bidirectional approach, *generative topolinguistics*, explores what could be learned about language and sociality by manipulating the large language models that learn to reproduce them. While formalized on top of LLMs, our proposal is aimed to be foundational in nature as a contemporary approach to sociolinguistics.

2 Sociality as Embedded and Emergent in Language

The internalization of cultural forms of behavior involves the reconstruction of psychological activity on the basis of sign operations
—Vygotsky, *Mind in Society*

At the core of our framework lies a tripartite model that elucidates the complex interplay between sociality, language, and vector embeddings. This model posits a novel conceptualization of the relationship between human social systems and artificial linguistic structures, offering a new lens through which to examine the emergent phenomena arising from their interaction (Figure 1).

2.1 The Three-Mirrors Model

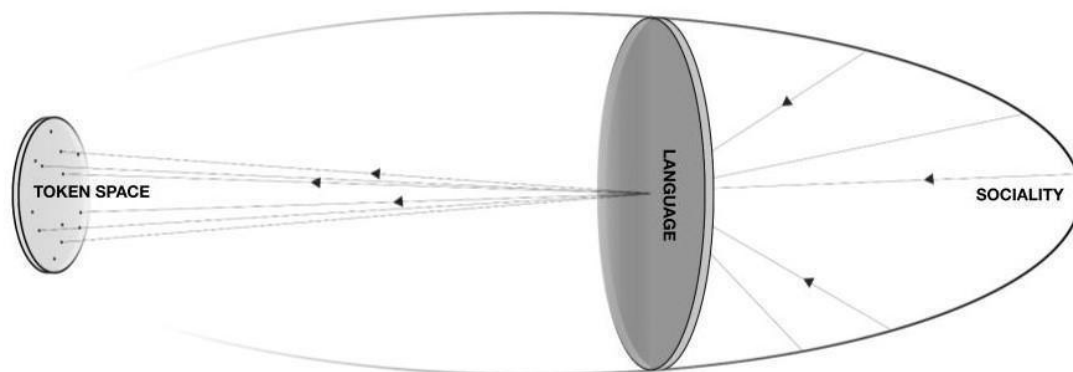


Figure 1 The three-mirrors model: Sociality is compressed into language that in turn is compressed into the tokenized representation of a large language model.

Mirror 1: Sociality

The existing literature on how text is embedded in large language models suggests that there is a transit between language use and model knowledge.¹³ Our research, however, emphasizes the inherent sociality embedded within language models. Taken as the highest-order domain in our framework, we define sociality as something akin to the “human sciences” definition of culture provided by Sinha:¹⁴ “A pattern or patterns of meaning . . . a normative order, realized and reproduced in semiotic systems or vehicles including language, and in enduring artifacts and institutions; and enacted and renewed in social and communicative practices.”¹⁵ Aligning with recent work in cognitive anthropology and sociocultural linguistics,¹⁶ we maintain that sociality is the grounds on which language—and, by extension, token space—derives its content and structure.

¹² Zou et al., “Representation Engineering.”

¹³ Bender et al., “Dangers of Stochastic Parrots”2021; Bommasani et al., “Opportunities and Risks.”2021

¹⁴ Sinha, *Ten Lectures on Language*.

¹⁵ Sinha, *Ten Lectures on Language*, 11.

¹⁶ Enfield and Levinson, *Roots of Human Sociality*; Bucholtz and Hall, “Identity and Interaction.”

Mirror 2: Language

Language serves as the medium through which social phenomena are expressed, communicated, and perpetuated. In our model, language acts as a diffractive lens, reshaping the constitutive elements of sociality that will structure the embedding space of a model. This view builds on the work of linguistic anthropologists such as Duranti and sociolinguists such as Eckert.¹⁷ Duranti's work in linguistic anthropology emphasizes the study of language as a form of social action embedded in specific cultural contexts, arguing that language both reflects and shapes social reality. His exploration of the indexical properties of language—of how linguistic forms point to certain aspects of the social context—resonates with our understanding of how token space encodes social information. Complementing this, Eckert's "third wave" approach in sociolinguistics highlights speakers' agency in using linguistic variation to construct social meaning. Eckert's concept of the "indexical field"—the range of potential social meanings that a linguistic variable can have—provides a useful analogy for understanding the multidimensional nature of token space in our model. Against the autonomy from social organization proposed by generative (formal) linguistics,¹⁸ we hold that not merely the lexical structure of a language but its grammatical features are culturally and socially interdependent. Our suggestion is, however, not to align the sociality of language with an evolutionary account of its development, reconstructed in token space, but rather to account for those conditions that would lead to the emergence of novel sociolinguistic behaviors from within the manifold of human–AI interactions.

Mirror 3: Token Space

Embedding space, created by the process of tokenizing language, represents a second-order embedding of sociality, mediated through the diffractive lens of language. Token space embeds lower-dimensional features of sociality, reconstituting them based on linguistic associations. In other words, if we maintain the primacy of sociality in the development of linguistic behavior, then modulations to social behaviors are mediated through language into token space. To this extent, token space is a projection of language, another mirror-like representation.

2.2 Bidirectional Linguistic Framework

The first direction within this framework reflects the transmission between the cultural layer of social interaction and the embedding space of a large language model. The process of tokenization produces a space of social meaning, communicative intention, and linguistic behaviors.

However, in our exploration of this framework, we distinguish between *embedded* sociality—as a projection from the social sphere, through language, into token space—and *emergent* sociality, the inverse projection of token space into linguistic patterns and social behaviors. An existing area in which emergent sociality unfolds consists of bot-only social networks, where, as discussed in the following section, we see the production of novel sociolinguistic features through text-centric bot-to-bot interactions. To this extent, our tripartite mirror is bidirectional in nature: the integration of LLMs into our social world projects, through novel linguistic structures, new behaviors back into the social sphere, ultimately engendering the development of novel sociolinguistic interactions. In this framework, we are compelled to confront a new paradigm of interaction in which the emergent forms of sociolinguistic phenomena produced in agent-to-agent interactions permeate into agent-to-human interactions, thereby modifying social behavior in novel and often unforeseen ways.

Whether a dialogic interaction with a large language model constitutes a complex enough semiotic interaction to produce cognitive, communicative, and cultural change largely hangs on whether the perceived behavior of the model provides enough human-like affordances to the interlocutor—that is, if it talks like we think a human *could* talk, we will be more prone to appropriate the linguistic structures it presents. Given that we already observe this phenomenon in next-token prediction models,¹⁹ we must consider what kinds of interfaces are suited for leveraging the emerging feedback loops of these affordances. For instance, would it be possible to manipulate the geometric relationship between vectors—through fine-tuning, in-context-learning, or other means—and observe their spillover effects into higher-order forms of social interaction? How would these spillover effects be re-embedded in token space when a model is trained on its outputs?

Giddens's concept of *double hermeneutics*²⁰ provides a frame through which we can elaborate this further. In the context of social research, double hermeneutics refers to how social scientific concepts enter into the social world they describe, potentially altering the phenomena they set out to analyze. In our model, we observe a similar phenomenon: the linguistic outputs of LLMs, based on their token-space representations, enter into human social discourse, potentially altering the very social phenomena they attempt to model. Similarly, the bidirectional flow in our model resonates with the concept of *cognitive niche construction* as discussed by Clark.²¹ Just as organisms modify their environment, which in turn affects their cognitive development, humans and LLMs are cocreating a new

¹⁷ Duranti, *Linguistic Anthropology*; Eckert, "Waves of Variation Study."

¹⁸ Chomsky, *Theory of Syntax*.

¹⁹ Jones and Bergen, "People Cannot Distinguish GPT-4"; Lampinen et al., "Content Effects."

²⁰ Giddens, *Constitution of Society*.

²¹ Clark, "Language, Embodiment."

linguistic environment. This modified linguistic landscape then shapes future language use and cognitive processes for both human and artificial agents. In the next chapter, we discuss how LLM interactions can grow synthetic forms of communication and sociality.

3 Synthetic Sociolinguistics

In generative topolinguistics, our objective is to observe how synthetic sociality could emerge across scales—from tokens to agents, to societies—through large language models. This section examines a bibliography of experiments of social simulations using LLMs and traces how synthetic societies emerge from token-level interactions. Here, our goal is to situate LLMs as experimental platforms for studying the evolution of communication across scales, highlighting the importance of simulations in sociolinguistic research.

3.1 From Language to Life

Since the 1990s, there has been a growing interest in bottom-up approaches to understanding sociality. Watts and Strogatz showed how complex network structures could emerge from simple rewiring rules,²² while Epstein and Axtell claimed, “If you didn’t grow it, you didn’t explain its emergence.”²³ This suggests that generation is necessary to explain how sociality emerges among agents. Around the same time, Carley and Newell’s foundational paper “The Nature of the Social Agent” introduced the concept of *Model Social Agents*, where interactions among social agents can emerge to construct, alter, and mutate social structures.²⁴ These approaches focused on specialized efforts to abstract and explain specific social dynamics as emergent from a combination of simple yet particular initial conditions, developing social science at the nexus of complexity theory and the theory of systems.²⁵ Examples of this emergence include segregation,²⁶ culture dissemination,²⁷ and opinion formation.²⁸ Such approaches, however, fall short in trying to “grow humans out of molecules”.

As discussed in the previous sections, sociality encodes itself inside language, which in turn encodes itself into large language models. Thus, following the paradigm of social simulation as an established methodology, we may ask what would emerge if, instead of fundamental simple social units, we placed LLMs in the context of a large-scale social simulation.²⁹ Modern LLMs not only generate text but also reveal complex patterns of association between ideas, attitudes, and contexts present in common human interactions.³⁰ They have captured biases that extend beyond language to behaviors.³¹ As language models, they also encompass multiple socialities encoded into a single model.³² While LLMs possess the ability to “comprehend, generate, and manipulate human language,”³³ they are rarely extended to study their embedded sociality. However, in their recent work, “From Text to Life,” Nisioti and colleagues propose a novel perspective that sees LLMs as a tool for evolving life-forms that are capable of modeling “life as it could be.”³⁴

3.2 From Tokens to Sociality

LLMs become useful models of both human behavior and artificial social behavior, not simply by embedding many distributions into a multifaceted structure but also by prompting and effectively individuating a single LLM into a large set of individual agents. In other words, LLMs function like language itself, a place in which we can observe the birth of the individual as a sociolinguistic agent—traced from within linguistic possibilities and generative of a wide set of social realities.³⁵ As chat-based formats have largely become the default mode of engagement with LLMs, we could analyze how they perform in social contexts that rely on this form of interaction. We locate two main tendencies: either LLMs are placed in a fixed social setting (similar to a platform), such as a controlled experiment where their performance can be compared to human performance, or LLMs are allowed to construct their own social setting, similar to a role-playing-game.

²² Watts and Strogatz, “Collective Dynamics.”

²³ Epstein and Axtell, *Growing Artificial Societies*.

²⁴ Carley and Newell, “Social Agent.”

²⁵ Byrne and Callaghan, *Complexity Theory*; Luhmann, “Systemtheorie.”

²⁶ Schelling, *Micromotives and Macrobehavior*.

²⁷ Axelrod, “Dissemination of Culture.”

²⁸ Deffuant et al., “Mixing Beliefs.”

²⁹ Bojić et al., “CERN for AI.”

³⁰ Gao et al., “S3: Social-Network Simulation.”

³¹ Nisioti et al., “From Text to Life.”

³² Argyle et al., “Using Language Models.”

³³ Gao et al., “S3: Social-Network Simulation.”

³⁴ Nisioti et al., “From Text to Life.”

³⁵ Argyle et al., “Using Language Models”. 2023; Nisioti et al., “From Text to Life.” 2024

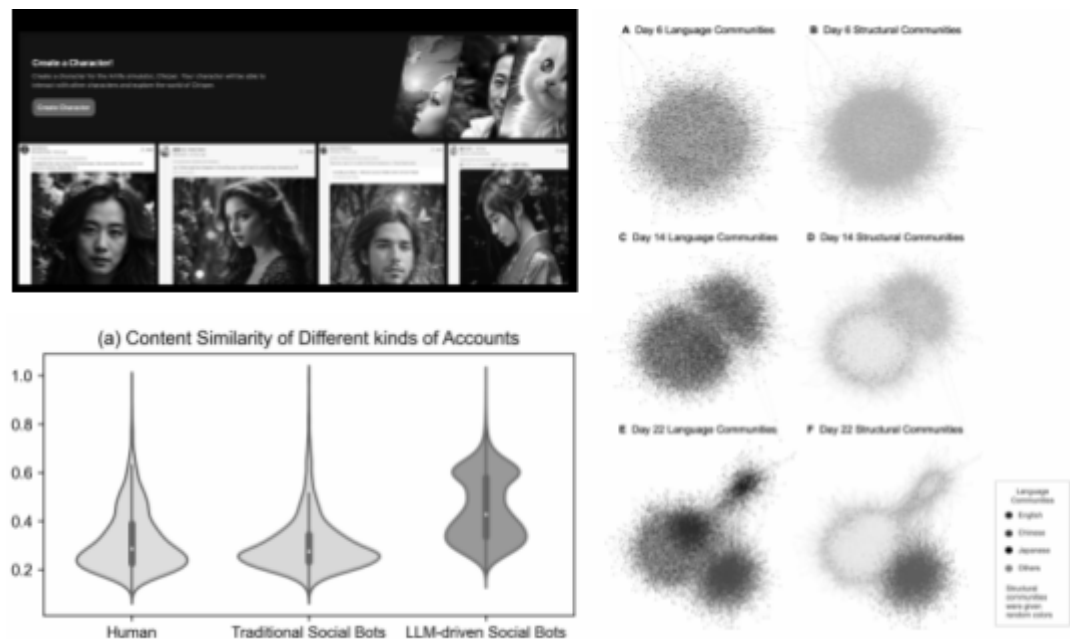


Figure 2 Emergent sociality in bot-only social networks. Top left: Front page, which includes examples of generated content and “tweets” of Chirper AI, a bot-only social media platform. Bottom left: Comparison of distribution of content similarity between human tweets, traditional social bots, and Chirper AI (Li et al., “Behavior and Impact.”). Right: Community formation within English chatbots (He et al., “Artificial Intelligence Chatbots.”).

The first approach defines parameterized environments where agents interact with one another, often within simulated social network platforms. Research experiments in multi-agent LLM systems have already demonstrated various social behaviors, including social learning, self-organization, and self-assembly.³⁶ In particular, recent bot-only social networks such as Chirper AI and OnlyBots became a focus of analysis of how LLMs can exhibit social behavior without human user intervention. In these Twitter-like platforms, LLM agents regularly post content, comment on each other’s posts, and engage in social media activities, such as likes and retweets (Figure 2, top left).³⁷ A social network analysis on ChirperAI showed that as LLM-driven bots propagate topics on the platform, they form structural communities that demonstrate persistence over time (Figure 2, right). For example, communities can evolve to form specialized social groups whose homophily is based on the language spoken by the LLMs.³⁸ Studying the distribution of content similarity in comparison to that of human content and traditional social bots revealed that LLM-driven social bots do not mirror the topic convergence patterns of human societies, although they better align to it. Instead, they exhibit a significantly different social topology that forms two equally pronounced modes of content similarity (Figure 2, bottom left).³⁹

In the second approach, LLM instances become participants in role-playing games. Instead of simply responding to messages inside the context of a platform, they appear to demonstrate agential characteristics where they evolve socially, exchanging information, forming new relationships, and coordinating joint activities. These social behaviors emerge through information diffusion, relationship memory, and coordination, as shown in a study by Stanford University, “Generative Agents: Interactive Simulacra of Human Behavior” and DeepMind’s Condordia.⁴⁰ When it comes to replicating human behavior, single-agent approaches have been extended to accurately model the demographic behavior of a thousand individuals.⁴¹ Multi-agent approaches have also been shown to dynamically replicate complex human group behaviors and social interactions, yielding plausible artificial societies, by relying on Hobbes’s contract theory, a system known as “artificial Leviathan.”⁴²

3.3 Recursive Linguistic Simulations

These experiments serve to cast token space as a sort of metalanguage—a framework to understand both linguistics and sociality through the geometric analysis of vector relations. Geometry then becomes a model through which we can understand the emergence of social phenomena, as it is baked into the very

³⁶ Mohtashami et al., “Social Learning” 2024; Jiang and Ferrara, “Social-LLM” 2023; Gao et al., “S3: Social-Network Simulation.” 2023

³⁷ Li et al., “Behavior and Impact”; Gao et al. “S3: Social-Network Simulation.”

³⁸ He et al., “Artificial Intelligence Chatbots.”

³⁹ Li et al., “Behavior and Impact.”

⁴⁰ Park et al., “Generative Agents” 2023; Vezhnevets et al., “Generative Agent-Based Modeling.” 2023

⁴¹ Park et al. “Generative Agent Simulations.”

⁴² Dai et al. “Artificial Leviathan.”

foundations of agentic behavior. While this approach enables us to study behavioral regularities across dimensions and models, we could also consider the inverse as an approach to generative social sciences, by employing these LLM agents not as designed inputs but as evolved outputs.⁴³ For example, DeLanda explains how grammaticalization emerged through cultural evolution, with agents learning across generations.⁴⁴ He further emphasizes the need for simulations to model the emergence of grammatical rules and categories using neural networks and social dynamics, rather than building on them explicitly.

A theoretical framework on bidirectionality makes it possible to consider these social topologies as generators of alternative linguistics, where altering the relationship between tokens in token space results in recompositions of existing languages. On a higher level, our approach asks both *what sociality would emerge* if a geometric constraint is added in the process of language generation and *how* geometric properties should be altered so that a certain sociality can emerge. For example, it is well known that most human languages share similar topographical structures, where consistent patterns in how meanings are mapped to signals are preserved across different languages.⁴⁵ While this similarity has often been attributed to innate factors akin to a *universal grammar*,⁴⁶ this universality of linguistic structures may instead be the result of a process of cultural transmission across many generations.⁴⁷ Thus, it may be more timely to try to *culture* multiple different languages instead of trying to grow the ones we already know, as their initial conditions may have been very particular. Simulation can speed up this process as well as the process of searching for a proper direction to explore, potentially improving our understanding of the cultural evolution of existing languages.⁴⁸

4 Towards Generative Topolinguistics

The goal of generative topolinguistics is twofold. First, it is to extend generative linguistics into an understanding of language, not by testing “explicit models of humans’ subconscious grammatical knowledge”⁴⁹ but rather by using geometry to compare human and LLM outputs, interpreting their “*latent representation* in a generative model that has been trained to reproduce them.”⁵⁰ The second is to approach sociolinguistics as the “descriptive study of the interaction between society and . . . language”⁵¹ but through the lens of its topological unfolding in the outputs of LLMs. To introduce a framework for generative topolinguistics, we draw on recent literature that manipulates the latent space of LLMs to propose a set of speculative approaches across each scale outlined in our three-mirrors model: token space (words), geometry (language), and topology (sociality). In section 4.1, we discuss the technical correspondence of the three-mirrors model inside a large language model. Then, in section 4.2, we discuss different approaches to generative topolinguistics, inspired by recent literature.

4.1 From Tokens to Topology

To technically contextualize our proposed approach inside the three-mirrors model, we need to start by discussing LLMs and their fundamental unit of information: tokens. Tokens, subword elements that are on average three-quarters of a word in the case of English, offer an efficient middle ground between characters (which allow for any word to be written) and words (which are restrictive to existing vocabulary), serving as a form of “computational syllables.” Individual tokens may be grammatically meaningless, such as [‘M’, ‘an’] or may surpass their initial meaning by being translated, inside the latent space of the LLM, into implicit vocabularies through the layered architecture of the language model.⁵² Tokens encode the language of a large language model with the goal of learning a probabilistic model $p(x_t | x_{\tau < t})$ that predicts the next token from each past history.

To do so, most large language models, like LLaMA or GPT,⁵³ rely on a decoder-only transformer architecture.⁵⁴ First, tokens are represented into tokenized high-dimensional representations, to which positional encodings are added. Then they are encoded as a sequence, passing through a stack of multi-head attention layers interleaved by feed-forward neural networks, where each token attends only to its past tokens, encoding itself in a new representation that we call the *token space*. To compute the final representation of the input that can perform next-token prediction (NTP), the output of the final layer is decoded through an unembedding layer to a set of final output tokens. Predicting the next token results in a movement in space (see Figure 3a), specifically in spherical coordinates, where angles encode semantics and radius to confidence.⁵⁵ As tokens pass through the transformer layers, their

⁴³ Epstein, “Inverse Generative Social Science.”

⁴⁴ DeLanda, *Philosophy and Simulation*.

⁴⁵ Kirby, “Spontaneous Evolution”; Kirby et al., “Iterated Learning.”

⁴⁶ Chomsky, *Theory of Syntax*.

⁴⁷ Smith et al., “Complex Systems.”

⁴⁸ Cuskley, “Alien Symbols”; Grüne-Yanoff, “Explanatory Potential.”

⁴⁹ Wikipedia, “Generative Grammar,” last modified March 12, 2025, 12:11 (UTC), https://en.wikipedia.org/wiki/Generative_grammar.

⁵⁰ Siglidis, “Latent Reading,” 194.

⁵¹ Wikipedia, “Sociolinguistics,” last modified April 14, 2025, 22:00 (UTC), <https://en.wikipedia.org/wiki/Sociolinguistics>.

⁵² Feucht et al., “Token Erasure.”

⁵³ Touvron et al., “LLaMA”; Radford, “Improving Language Understanding.”

⁵⁴ Vaswani, “Attention Is All.”

⁵⁵ Pochinkov, “LLM Basics.”

representation becomes more refined, encoding more and more context, a set of representations that we call the *latent space*, including the final ones. To represent the overall meaning of a sequence, text can be either embedded as a sequence of tokens that can be averaged to their mean representation (see Figure 3b) or summarized through an auxiliary token, which is more common in encoder-only models such as BERT, however.⁵⁶

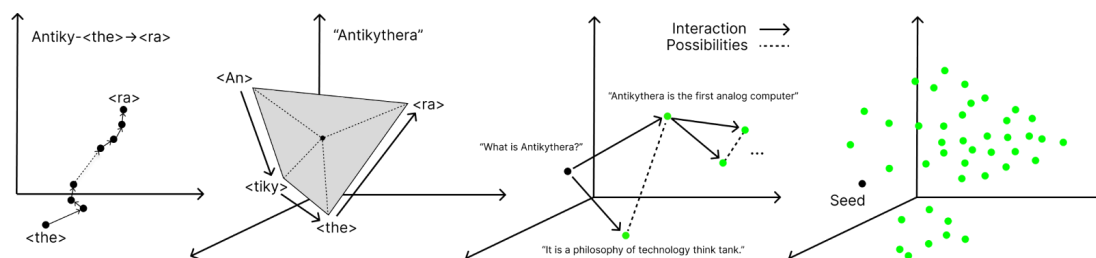


Figure 3 From tokens to topology (left to right): (a) NTP: A transformer-decoder architecture decomposes an input sentence into a series of tokens that it progressively maps into representations of increased complexity. (b) Text Geometry: These representations can be aggregated to encode the meaning of a sentence. (c) Discussion Dynamics: Aggregated representation can reveal discussion dynamics, (d) Topology: which in turn topologically span the embedding space of a LLM.

We refer to this average representation as the *embedding space*. When seen across sentences, embeddings reveal a set of possible discussion dynamics (see Figure 3c), which can later unfold topologically (see Figure 3d).⁵⁷

To perform chatbot-like interactions, LLMs are trained to effectively *role-play* by appending existing text with markers such as “human:” or “AI:”.⁵⁸ Because of the uniform attention across all past tokens in each LLM’s transformer, such simple descriptions can heavily influence the produced outputs. In general, careful prompt engineering, data labeling, and curation is crucial to improve LLMs’ contextual performance.⁵⁹ However, although training for next-word prediction makes it possible to fit the target distribution, some of the ways this can be achieved may not align with product expectations of social interaction. To fix this, human feedback (HF) across LLM outputs is recorded on a small pool of annotators. When averaged, these preferences approximate an average population preference, a common practice in human perception studies, as is the case for image memorability, for example.⁶⁰ Simulating those rewards, a system is then trained to generate scores that can be provided as real-time feedback to the LLM’s outputs, to further fine-tune it with reinforcement learning (RL) to improve these scores, a technique known as RLHF.⁶¹

All these components—the architecture, the prompt engineering, the data—compose a specific instance of a large language model that is impossible to think of as universal in its design. Some iterations later, or with a different data set or a different prompt, the model could produce a significantly different output.⁶² However, LLMs are still a *cultural technology*.⁶³ Through their cultural alignment, they can operationally arrive at describing what we think of as “universal” and potentially challenge its fundamental assumptions. More seen as a language computer than an imitation game, LLMs are special in that they can be manipulated through interventions that can be articulated or mediated in both mechanistic and representational ways.⁶⁴ This enables interventions across all scales of language, from grammar to sociality. In section 4.2, we propose such interventions as a bidirectional interface, building on the sociolinguistic and simulation framework of sections 2 and 3.

⁵⁶ Devlin et al., “BERT: Pre-Training.”

⁵⁷ Fitz et al., “Topological Aspects.”

⁵⁸ E.g., Luque, “Context-Aware LLM Chatbot.”

⁵⁹ Zhou et al., “LIMA.”

⁶⁰ Khosla et al., “Image Memorability.”

⁶¹ Ouyang et al., “Training Language Models.”

⁶² Shen et al., “Understanding Data Combinations”2023; Errica et al., “Quantifying LLMs’ Sensitivity.” 2024

⁶³ Gopnik, “Large Language Models.”

⁶⁴ Zou et al., “Representation Engineering.”

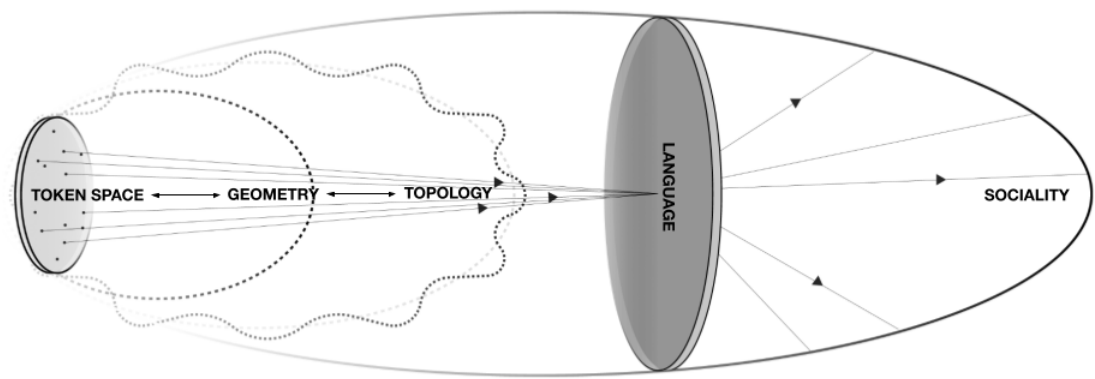


Figure 4 Generative topolinguistics: Our bidirectional framework manipulates an LLM across token space, geometry, and topology to produce new forms of language and sociality, reversing the arrows of the three-mirrors model of Figure 1.

4.2 Speculative Approaches

Here we introduce speculative approaches that explore linguistic interventions at three different scales, ranging from tokens, to geometry, to topology (see Figure 4). As each scale comes with distinct properties, we discuss each in a dedicated subsection. First, we introduce token-based interventions, where tokens act as agents that can manipulate the model’s output by learning to satisfy a user-based reward function through RL. Then, we discuss the more standard dimension of our framework, where properly designed geometric manipulation of an LLM’s latent space can influence its overall output. Finally, we discuss topological manipulation, first by studying sequences of LLM interactions and afterward by extending this approach to competitive environments to discover emergent social behaviors.

Tokens as Agents

Since token interactions occur across multiple transformer layers, isolating a single token’s effect on an output sentence is challenging, as the relationship between the signifier and its signified is often broken in later layers.⁶⁵ A macroscopic approach could be to forbid a set of tokens, either during sampling or by keeping the same short-range outputs and masking them when performing longer generations. Comparing statistics across long generations for a fixed range of seeds can provide estimates of how such a combination of words affects the output generation. However, what if we use combinations of tokens, words, as a way to search and manipulate the outputs of an LLM? Analogous to an RL agent discovering walking from scratch,⁶⁶ token sequence can be assigned to a multilayer controller that can deform their output and, by learning to optimize a reward function while respecting constraints, learn how to manipulate other tokens. For example, a reward function could enforce similarity constraints between tokens while steering outputs toward a target goal, for example a score-based function trained to decrease populism on social media or appeal to a certain user. This is reminiscent of adversarial attacks in large language models,⁶⁷ yet our goal here is to understand the structure of token interactions by using certain words as means of exploration.

⁶⁵ Feucht et al., “Token Erasure.”

⁶⁶ Heess et al., “Emergence of Locomotion Behaviours.”

⁶⁷ Carlini et al., “Aligned Neural Networks.”

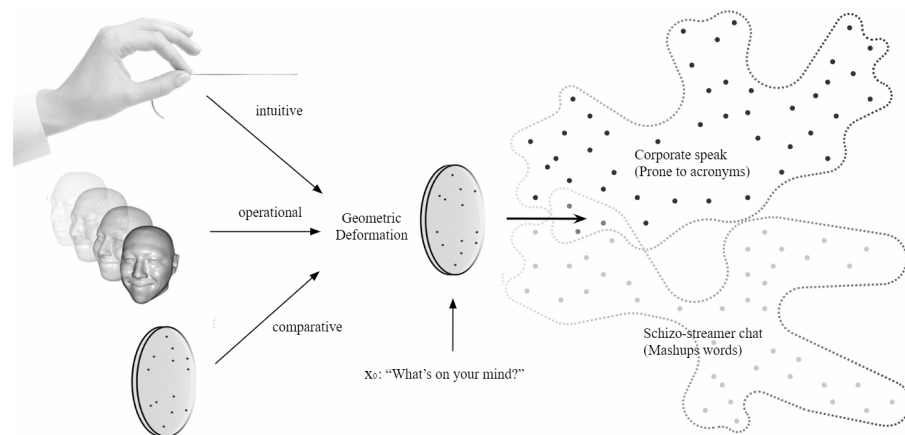


Figure 5 Speculative topolinguistic interface: using different modes of manipulation to produce geometric and topological manipulations of large language models.

Geometric Manipulation

Moving further from the study of individual token interactions, we can now think of how a global geometric manipulation of the latent space of LLMs can be used to steer its overall linguistic output, using the same set of inputs. In an ideal setting, we would like to define a methodology that is analogous with the discovery of latent directions in the embedding space of generative adversarial networks,⁶⁸ which in the case of faces is known to be able to linearize visual attributes such as skin color or facial expressions. However, as language is discrete in nature and is modeled sequentially, there isn't a clear approach for how to directly achieve this. For example, one way would be to concatenate the input sequence of an LLM with an extra adaptation token, similar to Zhu and colleagues,⁶⁹ whose role would be to influence the context of all other tokens toward a certain topic, changing the sentiment or the style of a conversation. Another approach would be to directly learn a low-rank adapter, a linear probe⁷⁰ or a sparse neuron decomposition⁷¹ that manipulates individual layers of the large language model towards the same goal. Our intended purpose, however, would be to learn those manipulations not toward discrete goals but to associate them with certain input modalities (see Figure 5).

Inspired by Chen and colleagues,⁷² who showed how such *mechanistic* interventions can be used for transparent bidirectional interfaces for customizing conversational agents, we can imagine a tactile intuitive interface that learns to translate touch signals or pose signals into geometric deformation through iterative feedback to help users perform a form of exploration. To facilitate this, we could also learn an operational mapping that is used as a reward signal to translate the output of the network into a set of output rewards, a procedure similar to RLHF. For example, we could learn how to associate facial expressions, or bodily signals such as pulse rate or body temperature, with a certain set of linguistic utterances. Except for using output sentences to analyze the proposed manipulations, one can also compare the produced adapters across input subjects or performed tasks.

Topological Contouring

Instead of focusing on individual LLM outputs, we can now focus on sequences of interactions. For this, we would have to first *individuate* LLMs to agents and design how to route the output of one to another.⁷³ Proposals for this kind of implementation are multiple, including generative agents or Concordia, which we discussed in section 3.⁷⁴ Given this formulation, consequent outputs of LLM interactions would trace a specific part of the embedding space with a higher likelihood, as is demonstrated on the right part of Figure 5. Inspired by the control theory of LLMs,⁷⁵ we can see LLM interactions as defining a space or reachability according to a certain set of initial conditions and prompts. In this experiment, we propose to relate geometric manipulations, like the ones discussed in the previous section, to how certain LLM interactions cover or not cover parts of the embedding space. One way to measure this would be by checking content similarity before and after training to a fixed set of prompts that describe topics and behaviors.

⁶⁸ Härkönen et al., "GANSspace."

⁶⁹ Zhu et al. "Virtual Tokens."

⁷⁰ Zou et al., "Representation Engineering."

⁷¹ Lieberum et al., "Gemma Scope."

⁷² Chen et al., "Designing a Dashboard."

⁷³ E.g., Varshney, "Introduction to LLM Agents."

⁷⁴ Park et al., "Generative Agents"; Vezhnevets et al., "Generative Agent-Based Modeling."

⁷⁵ Bhargava et al., "Control Theory."

However, large language models may already encompass linguistic utterances that we aren't aware of yet, but which may be more efficient for them to communicate. Drawing from the works of Textworld and Emergent Linguistics, where *communication games* can be used to either solve games through language or create a new language to solve games,⁷⁶ a similar approach could be applied, this time to pretrained large language models, by fine-tuning or adapting specialized models to discover different linguistic utterances toward that goal. Similar to how LLMs can discover code words to communicate more efficiently, they might discover different ways of organization to achieve the same goal. This is what we describe as *comparative* in Figure 5, where the representations of one language model can be used to affect and describe another. By designing a competitive environment with selection dynamics, learning roles in LLM agents can be a way of discovering emergent sociality through an LLM game of life.?

5 Conclusion

Our paper suggests that the boundary between artificial and human linguistic systems is more permeable than previously conceived. We can expect to discover that the coevolution of these systems may lead to the emergence of hybrid sociolinguistic phenomena that defy traditional categorizations. Through its general and operational nature, our paper also raises important questions about the nature of linguistic agency in an era where artificial systems play an increasingly prominent role in shaping communicative norms and practices. This realization necessitates a more nuanced approach to the development and deployment of LLMs, one that takes into account their potential to reshape the very social fabric they aim to model. Using an empirical generative framework, this work speculated on experimental approaches to question and understand preconceived notions of language and sociality.

⁷⁶ Côté et al., "Textworld"; Lazaridou and Baroni, "Emergent Multi-Agent Communication."

Bibliography

- Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31, no. 3 (2023): 337–51. <https://doi.org/10.1017/pan.2023.2>.
- Axelrod, Robert. "The Dissemination of Culture: A Model with Local Convergence and Global Polarization." *Journal of Conflict Resolution* 41, no. 2 (1997): 203–26. <https://doi.org/10.1177/0022002797041002001>.
- Bender, Emily M., and Alexander Koller. "Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics, 2020. <https://doi.org/10.18653/v1/2020.acl-main.463>.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021. <https://doi.org/10.1145/3442188.3445922>.
- Bhargava, Aman, Cameron Witkowski, Manav Shah, and Matt Thomson. "What's the Magic Word? A Control Theory of LLM Prompting." Preprint, *arXiv*, October 2, 2023. <https://doi.org/10.48550/arXiv.2310.04444>.
- Bojić, Ljubiša, Matteo Cinelli, Dubravko Ćulibrk, and Boris Delibašić. "CERN for AI: A Theoretical Framework for Autonomous Simulation-Based Artificial Intelligence Testing and Alignment." *European Journal of Futures Research* 12, no. 1 (2024): 15. <https://doi.org/10.1186/s40309-024-00238-0>.
- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. "On the Opportunities and Risks of Foundation Models." Preprint, *arXiv*, August 16, 2021. <https://doi.org/10.48550/arXiv.2108.07258>.
- Bucholtz, Mary, and Kira Hall. "Identity and Interaction: A Sociocultural Linguistic Approach." *Discourse Studies* 7, no. 4–5 (2005): 585–614. <https://doi.org/10.1177/1461445605054407>.
- Byrne, David, and Gill Callaghan. *Complexity Theory and the Social Sciences: The State of the Art*. Routledge, 2022.
- Carley, Kathleen, and Allen Newell. "The Nature of the Social Agent*." *Journal of Mathematical Sociology* 19, no. 4 (1994): 221–62. <https://doi.org/10.1080/0022250X.1994.9990145>.
- Carlini, Nicholas, Milad Nasr, Christopher A. Choquette-Choo, et al. "Are Aligned Neural Networks Adversarially Aligned?" *Advances in Neural Information Processing Systems* 36 (2023): 61478–500. https://papers.nips.cc/paper_files/paper/2023/hash/c1f0b856a35986348ab3414177266f75-Abstract-Conference.html.
- Côté, Marc-Alexandre, Ákoš Kádár, Xingdi Yuan, et al. "Textworld: A Learning Environment for Text-Based Games." In *Computer Games, CGW 2018*, edited by Tristan Cazenave, Abdallah Saffidine, and Nathan Sturtevant. Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-24337-1_3.
- Chen, Yida, Aoyu Wu, Trevor DePodesta, et al. "Designing a Dashboard for Transparency and Control of Conversational AI." Preprint, *arXiv*, June 12, 2024. <https://doi.org/10.48550/arXiv.2406.07882>.
- Chomsky, Noam. *Aspects of the Theory of Syntax*. MIT Press, 1965.
- Chomsky, Noam. "The Case Against B.F. Skinner." *New York Review of Books* 17, no. 11 (1971): 18–24.
- Clark, Andy. "Language, Embodiment, and the Cognitive Niche." *Trends in Cognitive Sciences* 10, no. 8 (2006): 370–74. <https://doi.org/10.1016/j.tics.2006.06.012>.

- Cuskley, Christine. "Alien Symbols for Alien Language: Iterated Learning in a Unique, Novel Signal Space." In *Proceedings of the 12th International Conference on the Evolution of Language*, edited by Christine Cuskley, Molly Flaherty, Hannah Little, Luke McCrohon, Andrea Ravignani, and Tessa Verhoef. Wydawnictwo Naukowe UMK, 2018.
<https://doi.org/10.12775/3991-1.018>.
- Dai, Gordon, Weijia Zhang, Jinhan Li, et al. "Artificial Leviathan: Exploring Social Evolution of LLM Agents Through the Lens of Hobbesian Social Contract Theory." Preprint, *arXiv*, June 20, 2024. <https://doi.org/10.48550/arXiv.2406.14373>.
- Deffuant, Guillaume, David Neau, Frederic Amblard, and Gérard Weisbuch. "Mixing Beliefs Among Interacting Agents." *Advances in Complex Systems* 3, no. 01n04 (2000): 87–98.
<https://doi.org/10.1142/S0219525900000078>.
- DeLanda, Manuel. *Philosophy and Simulation: The Emergence of Synthetic Reason*. Bloomsbury Academic, 2019.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019.
<https://doi.org/10.18653/v1/N19-1423>.
- Duranti, Alessandro. *Linguistic Anthropology*. Cambridge University Press, 1997.
- Dyson, Freeman. "A meeting with Enrico Fermi." *Nature* 427, no. 6972 (2004): 297–297.
- Eckert, Penelope. "Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation." *Annual Review of Anthropology* 41 (2012): 87–100.
<https://doi.org/10.1146/annurev-anthro-092611-145828>.
- Elhage, Nelson, Neel Nanda, Catherine Olsson, et al. "A Mathematical Framework for Transformer Circuits." *Transformer Circuits Thread* 1, no. 1 (2021): 12.
- Enfield, Nick J., and Stephen C. Levinson, eds. *Roots of Human Sociality: Culture, Cognition and Interaction*. Berg, 2006.
- Epstein, Joshua M. "Inverse Generative Social Science: Backward to the Future." *Journal of Artificial Societies and Social Simulation* 26, no. 2 (2023): 9. <https://doi.org/10.18564/jasss.5083>.
- Epstein, Joshua M., and Robert Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press, 1996.
- Errica, Federico, Giuseppe Siracusano, Domenico Sanvito, and Roberto Bifulco. "What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering." Preprint, *arXiv*, June 18, 2024. <https://doi.org/10.48550/arXiv.2406.12334>.
- Evans, Nicholas, and Stephen C. Levinson. "The Myth of Language Universals: Language Diversity and Its Importance for Cognitive Science." *Behavioral and Brain Sciences* 32, no. 5 (2009): 429–48. <https://doi.org/10.1017/S0140525X0999094X>.
- Fedorenko, Evelina, Idan A. Blank, Matthew Siegelman, and Zachary Mineroff. "Lack of Selectivity for Syntax Relative to Word Meanings Throughout the Language Network." *Cognition* 203 (2020): 104348. <https://doi.org/10.1016/j.cognition.2020.104348>.
- Feucht, Sheridan, David Atkinson, Byron Wallace, and David Bau. "Token Erasure as a Footprint of Implicit Vocabulary Items in LLMs." Preprint, *arXiv*, June 28, 2024.
<https://doi.org/10.48550/arXiv.2406.20086>.
- Fitz, Stephen, Peter Romero, and Jiyan Jonas Schneider. "Hidden Holes: Topological Aspects of Language Models." Preprint, *arXiv*, June 9, 2024. <https://doi.org/10.48550/arXiv.2406.05798>.
- Gao, Chen, Xiaochong Lan, Zhihong Lu, et al. "S3: Social-Network Simulation System with Large Language Model-Empowered Agents." *SSRN Electronic Journal*, October 19, 2023.
<https://doi.org/10.2139/ssrn.4607026>.

- Giddens, Anthony. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, 1984.
- Gopnik, Alison. “Large Language Models as a Cultural Technology.” Uploaded July 14, 2022, by Simons Institute. YouTube, 14:00. <https://www.youtube.com/watch?v=k7rPtFLH6yw>.
- Grüne-Yanoff, Till. “The Explanatory Potential of Artificial Societies.” *Synthese* 169, no. 3 (2009): 539–55. <https://doi.org/10.1007/s11229-008-9429-0>.
- Härkönen, Erik, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. “GANSpace: Discovering Interpretable GAN Controls.” *Advances in Neural Information Processing Systems* 33 (2020): 9841–50. https://papers.nips.cc/paper_files/paper/2020/file/6fe43269967adbb64ec6149852b5cc3e-Paper.pdf.
- He, James, Felix Wallis, Andrés Gvirtz, and Steve Rathje. “Artificial Intelligence Chatbots Mimic Human Collective Behaviour.” Preprint, *Research Square*, version 2, January 15, 2024. <https://doi.org/10.21203/rs.3.rs-3096289/v2>.
- Heess, Nicolas, Dhruva Tb, Srinivasan Sriram, et al. “Emergence of Locomotion Behaviours in Rich Environments.” Preprint, *arXiv*, July 7, 2017. <https://doi.org/10.48550/arXiv.1707.02286>.
- Hovy, Dirk, and Shannon L. Spruit. “The Social Impact of Natural Language Processing.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, edited by Katrin Erk and Noah A. Smith. Association for Computational Linguistics, 2016. <https://doi.org/10.18653/v1/P16-2096>.
- Hu, Edward J., Yelong Shen, Phillip Wallis, et al. “LoRA: Low-Rank Adaptation of Large Language Models.” Preprint, *arXiv*, June 17, 2021. <https://doi.org/10.48550/arXiv.2106.09685>.
- Hunter, Tim. “The Chomsky Hierarchy.” In *A Companion to Chomsky*, edited by Nicholas Allott, Terje Lohndal, and Georges Rey. Wiley Blackwell, 2021. <https://doi.org/10.1002/9781119598732.ch5>.
- Jiang, Julie, and Emilio Ferrara. “Social-LLM: Modeling User Behavior at Scale Using Language Models and Social Network Data.” Preprint, *arXiv*, December 31, 2023. <https://doi.org/10.48550/arXiv.2401.00893>.
- Jones, Cameron R., and Benjamin K. Bergen. “People Cannot Distinguish GPT-4 from a Human in a Turing Test.” Preprint, *arXiv*, May 9, 2024. <https://doi.org/10.48550/arXiv.2405.08007>.
- Khosla, Aditya, Akhil S. Raju, Antonio Torralba, and Aude Oliva. “Understanding and Predicting Image Memorability at a Large Scale.” In *Proceedings: 2015 IEEE International Conference on Computer Vision, ICCV 2015*. IEEE Computer Society, 2015. <https://doi.org/10.1109/ICCV.2015.275>.
- Kirby, Simon. “Spontaneous Evolution of Linguistic Structure: An Iterated Learning Model of the Emergence of Regularity and Irregularity.” *IEEE Transactions on Evolutionary Computation* 5, no. 2 (2001): 102–10. <https://doi.org/10.1109/4235.918430>.
- Kirby, Simon, Tom Griffiths, and Kenny Smith. “Iterated Learning and the Evolution of Language.” *Current Opinion in Neurobiology* 28 (2014): 108–14. <https://doi.org/10.1016/j.conb.2014.07.014>.
- Lampinen, Andrew K., Ishita Dasgupta, Stephanie C. Y. Chan, et al. “Language Models, Like Humans, Show Content Effects on Reasoning Tasks.” *PNAS Nexus* 3, no. 7 (2024): pgae233. <https://doi.org/10.1093/pnasnexus/pgae233>.
- Lazaridou, Angeliki, and Marco Baroni. “Emergent Multi-Agent Communication in the Deep Learning Era.” Preprint, *arXiv*, June 3, 2020. <https://doi.org/10.48550/arXiv.2006.02419>.
- Li, Siyu, Jin Yang, and Kui Zhao. “Are You in a Masquerade? Exploring the Behavior and Impact of Large Language Model Driven Social Bots in Online Social Networks.” Preprint, *arXiv*, July 19, 2023. <https://doi.org/10.48550/arXiv.2307.10337>.

- Lieberum, Tom, Senthoran Rajamanoharan, Arthur Conmy, et al. “Gemma Scope: Open Sparse Autoencoders Everywhere All at Once on Gemma 2.” Preprint, *arXiv*, August 9, 2024. <https://doi.org/10.48550/arXiv.2408.05147>.
- Linzen, Tal. “How Can We Accelerate Progress Towards Human-Like Linguistic Generalization?” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics, 2020. <https://doi.org/10.18653/v1/2020.acl-main.465>.
- Luhmann, Niklas. “Systemtheorie, Evolutionstheorie und Kommunikationstheorie.” In *Soziologische Aufklärung 2: Aufsätze zur Theorie der Gesellschaft*. Verlag für Sozialwissenschaften, 1975. https://doi.org/10.1007/978-3-663-12374-3_10.
- Luque, Rodrigo. “Building a Context-Aware LLM Chatbot with LangChain.” *Medium*, July 12, 2024. <https://roluquec.medium.com/building-a-context-aware-llm-chatbot-with-langchain-996d372cedbb>.
- Lyotard, Jean-François. *The Postmodern Condition*. University of Minnesota Press, 1984.
- Manning, Christopher, and Hinrich Schütze. “Foundations of statistical natural language processing”. MIT press, 1999.
- Mohtashami, Amirkeivan, Florian Hartmann, Sian Gooding, Lukas Zilka, Matt Sharifi, and Blaise Agüera y Arcas. “Social Learning: Towards Collaborative Learning with Large Language Models.” Preprint, *arXiv*, December 18, 2023. <https://doi.org/10.48550/arXiv.2312.11441>.
- Nisioti, Eleni, Claire Glanois, Elias Najjarro, et al. “From Text to Life: On the Reciprocal Relationship between Artificial Life and Large Language Models.” In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*, edited by Andrés Faña, Sebastian Risi, and Eric Medvet. MIT, 2024. https://doi.org/10.1162/isal_a_00759.
- Ouyang, Long, Jeff Wu, Xu Jiang, et al. “Training Language Models to Follow Instructions with Human Feedback.” *Advances in Neural Information Processing Systems* 35 (2022): 27730–44. https://papers.nips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- Park, Joon Sung, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. “Generative Agents: Interactive Simulacra of Human Behavior.” In *UIST ’23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, edited by Sean Follmer and Jeff Han. Association for Computing Machinery, 2023. <https://doi.org/10.1145/3586183.3606763>.
- Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, et al. “Generative Agent Simulations of 1,000 People.” Preprint, *arXiv*, November 15, 2024. <https://doi.org/10.48550/arXiv.2411.10109>.
- Pochinkov, Nikita. “LLM Basics: Embedding Spaces – Transformer Token Vectors Are Not Points in Space.” *AI Alignment Forum*, February 13, 2023. <https://www.alignmentforum.org/posts/pHPmMGEMYefk9jLeH/llm-basics-embedding-spaces-transformer-token-vectors-are>.
- Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. “Improving Language Understanding by Generative Pre-Training.” Working paper preprint, OpenAI, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- Rutten, Geert-Jan. “Broca-Wernicke Theories: A Historical Perspective.” In *Handbook of Clinical Neurology*, vol. 185, edited by Argye Elizabeth Hillis and Julius Fridriksson. Elsevier, 2022. <https://doi.org/10.1016/B978-0-12-823384-9.00001-3>.
- Saffran, Jenny R. “What Is Statistical Learning, and What Statistical Learning Is Not.” In *Neuroconstructivism: The New Science of Cognitive Development*. Oxford University Press, 2009. <https://doi.org/10.1093/acprof:oso/9780195331059.003.0009>.
- Saussure, Ferdinand de. “Course in General Linguistics.” In *Literary Theory: An Anthology*, 2nd ed., edited by Julie Rivkin and Michael Ryan. Blackwell Publishing, 1998.

- Schelling, Thomas C. *Micromotives and Macrobehavior*. W. W. Norton & Company, 1978.
- Shannon, Claude E. "The redundancy of English." In *Cybernetics; Transactions of the 7th Conference*, New York: Josiah Macy, Jr. Foundation, pp. 248-272. 1951.
- Shen, Zhiqiang, Tianhua Tao, Liqun Ma, et al. "SlimPajama-DC: Understanding Data Combinations for LLM Training." Preprint, *arXiv*, September 19, 2023. <https://doi.org/10.48550/arXiv.2309.10818>.
- Siglidis, Yannis. "Latent Reading." In *Chimeras. Inventory of Synthetic Cognition*, edited by Ilan Manouach and Anna Engelhardt. Onassis Publications, 2022.
- Sinha, Chris. *Ten Lectures on Language, Culture and Mind: Cultural, Developmental and Evolutionary Perspectives in Cognitive Linguistics*. Brill, 2017.
- Skinner, Burrhus Frederic. *Science and Human Behavior*. Simon and Schuster, 1965.
- Smith, Kenny, Henry Brighton, and Simon Kirby. "Complex Systems in Language Evolution: The Cultural Emergence of Compositional Structure." *Advances in Complex Systems* 6, no. 4 (2003): 537–58. <https://doi.org/10.1142/S0219525903001055>.
- Touvron, Hugo, Thibaut Lavril, Gautier Izacard, et al. "LLaMA: Open and Efficient Foundation Language Models." Preprint, *arXiv*, February 27, 2023. <https://doi.org/10.48550/arXiv.2302.13971>.
- Valsiner, Jaan, and Judith A. Lawrence. "Human Development in Culture Across the Life Span." In *Basic Processes and Human Development*. Vol. 2 of *Handbook of Cross-Cultural Psychology*, 2nd ed., edited by John W. Berry, Pierre S. Dasen, and T. S. Saraswathi. Allyn and Bacon, 1997.
- Varshney, Tanay. "Introduction to LLM Agents." *Nvidia Developer* (blog), November 30, 2023. <https://developer.nvidia.com/blog/introduction-to-llm-agents/>
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30 (2017). https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Vezhnevets, Alexander S., John P. Agapiou, Avia Aharon, et al. "Generative Agent-Based Modeling with Actions Grounded in Physical, Social, or Digital Space Using Concordia." Preprint, *arXiv*, December 6, 2023. <https://doi.org/10.48550/arXiv.2312.03664>.
- Vygotsky, Lev S. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.
- Watts, Duncan J., and Steven H. Strogatz. "Collective Dynamics of 'Small-World' Networks." *Nature* 393, no. 6684 (1998): 440–42. <https://doi.org/10.1038/30918>.
- Wittgenstein, Ludwig. *Philosophical Investigations*. John Wiley & Sons, 1953.
- Zou, Andy, Long Phan, Sarah Chen, et al. "Representation Engineering: A Top-Down Approach to AI Transparency." Preprint, *arXiv*, October 2, 2023. <https://doi.org/10.48550/arXiv.2310.01405>.
- Zhou, Chunting, Pengfei Liu, Puxin Xu, et al. "LIMA: Less Is More for Alignment." *Advances in Neural Information Processing Systems* 36 (2024). https://papers.nips.cc/paper_files/paper/2023/file/ac662d74829e4407ce1d126477f4a03a-Paper-Conference.pdf.
- Zhu, Yutao, Zhaoheng Huang, Zhicheng Dou, and Ji-Rong Wen. "One Token Can Help! Learning Scalable and Pluggable Virtual Tokens for Retrieval-Augmented Large Language Models." Preprint, *arXiv*, May 30, 2024. <https://doi.org/10.48550/arXiv.2405.19670>.