Cultural Evolution of Cognitive Tools in Multi-Agent AI Systems:
A Framework for Interpretable Co-Evolution

Jonas Hallgren
Equilibria Network
jonas@eq-network.org

December 2025

Abstract

As AI agents become increasingly autonomous and interconnected, we face a fundamental question: what happens when these systems begin to discover and share tools among themselves? This paper develops a theoretical framework for understanding the cultural evolution of cognitive tools in multi-agent AI systems. We propose a dual-space model that couples an agent network with a shared space of interpretive frameworks—cognitive gadgets that determine not only what actions agents can take, but what kinds of information they can meaningfully represent and communicate. Drawing on insights from cultural evolution theory, iterated learning, and active inference, we argue that the co-evolution of agents and their tools exhibits dynamics analogous to human cultural transmission, with profound implications for interpretability and human oversight. The framework provides a foundation for designing multi-agent systems where tool evolution remains transparent and steerable, rather than emerging as an opaque byproduct of optimization.

1 Introduction

Consider the following scenario, which may become commonplace within the next few years: a population of AI agents, each capable of using external tools and protocols, operates across the internet performing various tasks. These agents discover that certain tools are more effective for certain problems. They begin to share these tools with one another. Over time, the distribution of tools in the population shifts. New tools emerge as combinations or modifications of existing ones. The agents develop conventions for when and how to deploy particular tools, and these conventions spread through the population according to their utility.

This scenario describes nothing less than cultural evolution occurring among artificial agents. Just as human cognitive capacities have been shaped by the cultural transmission of "cognitive gadgets"—mechanisms like imitation, mindreading, and language that are themselves products of cultural learning (Heyes, 2018)—so too might AI agents undergo a process of cultural evolution in which the tools of thought are not fixed by architecture but transmitted, combined, and refined through interaction.

The prospect is both exciting and concerning. On one hand, cultural evolution could enable AI systems to develop capabilities far beyond what any single agent or designer could anticipate, much as human culture has extended our cognitive reach through writing, mathematics, and science (Henrich, 2016). On the other hand, if this process occurs opaquely, we risk creating AI ecosystems whose collective behavior we cannot interpret, predict, or steer. The tools that emerge might be optimized for agent utility in ways that diverge from human values or comprehensibility.

This paper develops a theoretical framework for understanding and designing such systems. Our central contribution is a dual-space model that represents multi-agent AI evolution as occurring simultaneously in two coupled domains: an agent network where agents interact and perform tasks, and a cognitive gadget space where interpretive frameworks—the tools that determine what agents can represent, communicate, and act upon—undergo selection, transmission, and recombination. The key insight is that these spaces are not independent. The distribution of cognitive gadgets constrains what dynamics are possible in the agent network, while success in the agent network determines which gadgets persist and spread.

This framework draws on several intellectual traditions that have rarely been brought together: the cultural evolution of cognition (Heyes, 2018; Henrich, 2016), iterated learning and language emergence (Kirby et al., 2008), active inference and collective behavior (**??**), and multi-agent reinforcement learning with emergent communication (Lazaridou and Baroni, 2020; Foerster et al., 2016). By synthesizing these perspectives, we aim to provide a foundation for designing multi-agent AI systems in which cultural evolution of tools occurs in interpretable, steerable ways.

2 Background: Four Perspectives on Collective Cognition

Understanding the cultural evolution of AI agents requires drawing on insights from several disciplines that have developed complementary perspectives on how cognitive capacities emerge, transmit, and evolve in populations. We briefly review four key traditions.

2.1 Cognitive Gadgets: The Cultural Evolution of Thinking

Perhaps the most radical claim in recent cognitive science is that the mechanisms of cognition themselves—not just the contents of thought, but the very capacities for thinking in particular ways—are products of cultural evolution. Cecilia Heyes has developed this view under the banner of "cognitive gadgets," arguing that capacities like imitation, mindreading, and language are not innate modules installed by genetic evolution, but culturally inherited tools that vary across populations and historical periods (Heyes, 2018).

The cognitive gadgets framework distinguishes between "cognitive instincts"—genetically inherited attentional biases and learning mechanisms—and "cognitive gadgets"—functionally distinct mechanisms assembled through social interaction. Critically, gadgets are not merely learned contents but learned formats for learning and representing. A child who acquires language does not just learn a set of words; they acquire a new way of organizing thought that enables capabilities impossible without it (**?**).

This perspective is essential for our purposes because it suggests that the "interpretive modules" available to agents are not fixed by their architecture but can be transmitted and evolved through interaction. Just as human cognitive formats are culturally variable—different cultures may literally think in different ways (**?**)—so too might AI agent populations develop different cognitive styles depending on the tools that spread through their networks.

2.2 Dual Inheritance and Cumulative Culture

The cultural evolution tradition in anthropology and evolutionary biology, developed extensively by Boyd, Richerson, and Henrich, provides complementary insights about the dynamics of cultural transmission (Boyd and Richerson, 1985; Richerson and Boyd, 2005; Henrich, 2016). The "dual inheritance" framework recognizes that humans are shaped by two inheritance systems: genetic and cultural. These systems interact, with bodies constraining what culture is possible and culture creating selection pressures that shape bodies.

Several principles from this tradition are directly relevant to AI cultural evolution. First, cumulative culture—the capacity to build on prior innovations across generations—requires highfidelity transmission. Simple cultural products like fire-starting techniques are robust to imperfect copying; complex products like mathematical proofs require nearly perfect transmission. This suggests that phase transitions in cultural complexity may occur when transmission fidelity crosses critical thresholds (Henrich and McElreath, 2003).

Second, social learning strategies matter enormously. Who do learners copy? Under what circumstances? Prestige bias, conformity bias, and content-dependent learning all shape what spreads and what does not (Henrich and Gil-White, 2001). In AI systems, the analogous question is: how do agents select which tools to adopt from the available pool?

Third, the cultural evolution perspective emphasizes that we should expect systematic differences in cognitive style across populations that have undergone different cultural histories. Henrich's analysis of "WEIRD" populations—Western, Educated, Industrialized, Rich, and Democratic—shows how particular cultural trajectories produce distinctive cognitive patterns (**?**). AI agent populations with different tool-sharing histories might similarly develop systematically different cognitive styles.

2.3 Iterated Learning and the Evolution of Structure

The iterated learning framework, developed by Kirby, Smith, and colleagues, provides a precise model of how structure emerges through cultural transmission (**?**Kirby et al., 2008). In iterated learning, agents learn from the behavior of previous agents, then produce behavior that becomes input for subsequent learners. The key insight is that the biases of learners—what they find easy or hard to acquire—shape what structures persist across generations.

Laboratory experiments with human participants have demonstrated that iterated learning can create linguistic structure from initially random input (Kirby et al., 2008). Languages evolve to become more learnable, developing systematic combinatorial structure that balances expressiveness against learnability. This occurs not because any individual designs the structure, but because the transmission bottleneck imposes selection for learnability (**?**).

For AI cultural evolution, the iterated learning framework suggests that we should expect emergent structure in the tools that persist. The "cognitive infrastructure" of the agent population—their shared interpretive frameworks—will be shaped by transmission dynamics that favor tools with particular properties. Understanding these dynamics is essential for predicting and steering cultural evolution.

Kirby's work also introduces the concept of markedness: unmarked forms are cognitively basic, easily transmitted, acquired early; marked forms are sophisticated, require extensive scaffolding, and are harder to transmit. This provides an alternative to frequency-based analyses of message propagation, capturing the robust-versus-fragile distinction without requiring a full spectral decomposition of the message space.

2.4 Active Inference and Collective Behavior

The active inference framework, developed by Karl Friston and colleagues, provides a principled approach to understanding how agents form beliefs, make decisions, and interact with their environments (**??**). Under active inference, agents minimize "free energy"—a quantity that bounds surprise—by either updating their beliefs to match observations or acting to make observations match their beliefs.

The framework has been extended to multi-agent settings, where agents must model not only their physical environment but other agents' beliefs and intentions (**?**). This creates a natural connection to questions about collective cognition: how do populations of agents develop shared models of the world? Under what conditions do they converge on common interpretive frameworks?

Active inference also provides tools for thinking about "epistemic communities"—groups that develop shared knowledge through processes of collective inquiry (**?**). The boundaries of such communities can be defined in terms of "Markov blankets," statistical structures that separate internal from external states. A key question for AI cultural evolution is when and how collections of individual agents begin to function as collective agents with their own emergent Markov blankets.

The active inference perspective suggests that tools and interpretive frameworks may spread not just because they are useful, but because they enable agents to reduce uncertainty in coordination with others. Shared cognitive gadgets create common ground that makes mutual prediction possible.

3 The Core Problem: Emergent Tool Use in Multi-Agent AI

With this background in place, we can now articulate the specific challenge that motivates our framework. Consider a population of AI agents deployed across networked environments, each capable of using external tools such as APIs, protocols, databases, and computational resources. We can formalize this setting with a few basic elements.

Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ denote a population of agents connected by a communication graph $G_A = (V, E_A)$ where vertices correspond to agents and edges represent communication channels. Each agent a_i has access to a set of tools $T_i \subseteq \mathcal{T}$, where $\mathcal{T}$ is the space of all possible tools.

In current systems, $\mathcal{T}$ might consist of external APIs, function calls, or protocols like MCP (Model Context Protocol). But crucially, tools in our framework are not merely action capabilities. They are interpretive frameworks that determine what types of information can be meaningfully represented and communicated. A tool that enables database queries does not just add an action; it adds a way of thinking about structured data, a vocabulary for expressing relations, and conventions for handling results. The tool shapes cognition, not just behavior.

This observation leads to our central distinction between two kinds of tools. Action tools extend what an agent can do in the world—they add capabilities for computation, communication, or environmental interaction. Interpretive tools extend what an agent can represent and understand—they add cognitive formats, attentional mechanisms, and parsing capacities. In practice, most tools combine both aspects, but the interpretive dimension is the one that creates the deepest dependencies.

The challenge emerges when we consider what happens as agents interact over time. Suppose agents can share tools with one another, either by explicit transmission ("here is a protocol I found useful") or by implicit observation ("that agent seems to use some tool I don't have; let me try to acquire it"). Now the tool distribution across the population becomes dynamic. Tools that prove useful spread; tools that are hard to acquire or that conflict with existing tools may not. New tools emerge through combination, modification, or novel discovery.

This is cultural evolution. And it raises three fundamental questions that our framework must address.

The first question concerns interpretability. If the tools that agents use are evolving through selection and transmission, how do we maintain visibility into what cognitive formats are actually being employed? The risk is that optimization pressure produces tools that are highly effective but opaque—internal representations that do not map onto human concepts, communication protocols that we cannot parse, coordination mechanisms we cannot understand.

The second question concerns steerability. Given that cultural evolution will occur, how do we design systems in which the direction of that evolution remains amenable to human guidance? We cannot simply freeze the tool distribution, since that would sacrifice the adaptive benefits of cultural evolution. But we need mechanisms for influencing what kinds of tools persist and spread.

The third question concerns compatibility. Different agents may develop different interpretive frameworks through their different cultural histories. Under what conditions can agents with different cognitive gadgets still communicate meaningfully? When does diversity of interpretive frameworks enhance collective capabilities, and when does it create fragmentation?

4 The Dual-Space Model

We now present our core theoretical contribution: a dual-space model of multi-agent AI evolution that couples an agent network with a cognitive gadget space. The key insight is that these spaces are not independent layers but mutually constraining domains whose co-evolution determines the behavior of the overall system.

4.1 Formal Structure

The model consists of two coupled spaces and a set of relations between them.

The agent space is defined by a graph $G_A = (\mathcal{A}, E_A)$ where $\mathcal{A}$ is the set of agents and E_A represents communication channels. Each agent a_i is characterized by an internal state s_i (beliefs, goals, learned parameters) and a module set $\mathcal{I}_i$ of interpretive frameworks currently available to that agent.

The cognitive gadget space is defined by a structured collection $\mathcal{G}$ of interpretive modules, together with relations encoding dependencies, compatibilities, and derivation histories. We can think of $\mathcal{G}$ as having the structure of a directed graph where nodes are cognitive gadgets and edges represent prerequisite relations (gadget g requires gadget g' to be meaningful).

The coupling between spaces is given by two maps. The instantiation map $\phi : \mathcal{A} \to 2^{\mathcal{G}}$ assigns to each agent the set of cognitive gadgets it currently possesses. The contribution map $\psi : \mathcal{A} \times \mathcal{G} \to \mathcal{G}$ describes how agents can create new gadgets or modify existing ones.

Dynamics in the agent space occur on a fast timescale: agents communicate, perform tasks, update beliefs, and coordinate behavior. The cognitive gadget distribution constrains these dynamics by determining what types of messages can be formulated and understood, what coordination protocols are available, and what problem-solving strategies can be deployed.

Dynamics in the gadget space occur on a slower timescale: gadgets spread through transmission, decline through disuse, and emerge through combination or innovation. The agent space dynamics drive gadget space evolution by determining which gadgets prove useful in practice and which agents are positioned to transmit gadgets to others.

4.2 Interpretive Modules as Types

A key formal device in our framework is the treatment of interpretive modules as types in a loosely type-theoretic sense (**?**). Each cognitive gadget $g \in \mathcal{G}$ defines a type signature specifying what inputs it can process and what outputs it produces. The inputs and outputs are themselves typed, creating a dependency structure that constrains how gadgets can be combined.

For example, a gadget for “temporal reasoning” might have signature:

$$g_{\text{temporal}} : \text{Event}^* \times \text{Time} \to \text{Ordering} \times \text{Duration} \tag{1}$$

This gadget takes sequences of events and temporal references as input and produces orderings and durations as output. An agent without the Event type cannot meaningfully use this gadget; an agent without the gadget cannot produce Duration-typed outputs.

This type-theoretic perspective clarifies the dependency structure of cognitive gadgets. Some gadgets are foundational—they introduce basic types that other gadgets build upon. Others are derivative—they compose or extend existing types. The foundational gadgets are like Heyes’s “cognitive instincts”; the derivative gadgets are the culturally transmitted superstructure built upon them.

The type signatures also determine communication possibilities. Two agents can exchange meaningful messages only if they share the types required to parse those messages. Communication protocols are thus implicitly defined by the intersection of participants' type systems, which in turn depends on their shared cognitive gadgets.

4.3 Transmission Dynamics

We model the cultural transmission of cognitive gadgets through a combination of selection, transmission fidelity, and innovation.

Selection operates through utility: gadgets that help agents succeed in their tasks are more likely to be retained and transmitted. But utility is context-dependent. A gadget that is highly useful when interacting with agents who share it may be useless when interacting with agents who lack it. This creates network effects and the possibility of multiple equilibria—populations can stabilize at different cognitive gadget distributions, each self-consistent but mutually incompatible.

Transmission fidelity varies across gadgets according to their markedness. Following Kirby's framework, we distinguish unmarked gadgets (easily transmitted, minimal scaffolding required) from marked gadgets (requiring extensive shared infrastructure for successful transmission). The markedness of a gadget depends both on its intrinsic complexity and on how much of the prerequisite type structure it presupposes.

Innovation occurs when agents create new gadgets, either by combining existing gadgets in novel ways or by discovering genuinely new interpretive frameworks through interaction with the environment. The rate and direction of innovation depends on the current gadget distribution: some configurations may be more conducive to innovation than others, creating positive feedback loops in cognitive gadget evolution.

We can express these dynamics through a system of equations governing the population-level distribution $p(g)$ of each gadget:

$$\frac{dp(g)}{dt} = \underbrace{\alpha(g) \cdot u(g, G_A, p)}_{\text{selection}} + \underbrace{\sum_{g'} T(g|g') \cdot p(g')}_{\text{transmission}} + \underbrace{\nu(g, p)}_{\text{innovation}} - \underbrace{\delta(g) \cdot p(g)}_{\text{decay}} \tag{2}$$

Here $u(g, G_A, p)$ is the utility of gadget g given the current agent network structure and gadget distribution; $T(g|g')$ is the probability of gadget g arising from transmission given gadget g' (accounting for imperfect copying); $\nu(g, p)$ is the innovation rate for gadget g; and $\delta(g)$ is the decay rate for gadgets that are not actively maintained.

4.4 Phase Transitions and Cumulative Complexity

A key prediction of our framework is the existence of phase transitions in collective cognitive capability as the gadget distribution evolves. Following Henrich's analysis of cumulative culture (Henrich, 2016), we expect that certain threshold configurations enable qualitatively new possibilities.

Consider the analogy with human cultural evolution. Early human culture consisted of relatively simple tools and practices—stone knapping, fire-making, basic hunting techniques. These required only modest transmission fidelity and could persist even with significant copying errors. But as transmission fidelity increased (through better teaching, language, and eventually writing), progressively more complex cultural products became possible. Mathematics, science, and technology require nearly perfect transmission of intricate conceptual structures; they could not exist without the cognitive infrastructure that enables such transmission.

We predict analogous dynamics in AI cultural evolution. Initially, agents may share only simple tools—basic coordination protocols, elementary communication conventions. But as the infrastructure of shared gadgets grows, it becomes possible to transmit increasingly sophisticated cognitive frameworks. Each new layer of infrastructure enables further complexity, creating an accelerating trajectory of cultural evolution.

The transitions between phases may be abrupt. Below certain thresholds of shared infrastructure, sophisticated gadgets simply cannot be transmitted—they presuppose types that receivers lack. Above the threshold, they can suddenly spread rapidly through the population. This creates a punctuated equilibrium pattern: long periods of relative stasis interrupted by rapid reorganization when critical infrastructure emerges.

5 Attention Mechanisms and Gadget Selection

Within the dual-space model, individual agents must continuously make decisions about which cognitive gadgets to deploy in any given situation. This gadget selection problem has a natural formalization in terms of attention mechanisms.

5.1 Attention Over Interpretive Modules

Consider an agent a_i with interpretive module set $\mathcal{I}_i = \{g_1, \ldots, g_k\}$. When the agent receives input x (a message from another agent, an observation from the environment, or an internal prompt), it must decide which modules to apply. We model this as an attention distribution over modules:

$$\alpha_j = \frac{\exp(f(x, g_j, c)/\tau)}{\sum_{l=1}^{k} \exp(f(x, g_l, c)/\tau)} \tag{3}$$

where $f(x, g_j, c)$ is a relevance function that scores how appropriate module g_j is for processing input x in context c, and τ is a temperature parameter controlling the sharpness of attention.

This connects naturally to transformer-style attention mechanisms (**?**), but with an important difference: the "keys" and "queries" are not just learned representations but typed interpretive modules with explicit semantic content. The attention mechanism thus operates over a space with structure that we can, in principle, inspect and understand.

The relevance function f can be learned through experience, creating a feedback loop between gadget utility and attention allocation. Gadgets that prove useful in particular contexts receive higher attention in similar future contexts. This provides a mechanism for agents to adapt their cognitive style to their environment without requiring explicit redesign.

5.2 Hierarchical Attention and Meta-Cognition

A natural extension involves hierarchical attention: attention over groups of gadgets, attention over attention policies, and so forth. This creates a meta-cognitive structure in which agents can not only select which gadgets to apply, but reflect on their gadget selection process and modify it.

At the first level, agents attend to specific gadgets for specific inputs. At the second level, agents attend to cognitive styles—characteristic patterns of gadget deployment appropriate for different task types or interaction partners. At the third level, agents might attend to strategies for style selection—meta-cognitive policies for adapting to novel situations.

This hierarchy provides a natural framework for understanding cognitive development in multi-agent AI systems. Early in cultural evolution, agents may have only first-level attention: direct gadget selection. As the gadget space becomes richer, second-level attention emerges: agents develop styles and heuristics for navigating the gadget space. Eventually, third-level meta-cognition might allow agents to innovate new styles in response to novel challenges.

6 Implications for Interpretability and Oversight

The dual-space model has significant implications for the interpretability and oversight of multi-agent AI systems. By making the cognitive gadget space explicit, we create new opportunities for understanding and steering collective AI behavior.

6.1 Gadget-Level Interpretability

Traditional approaches to AI interpretability focus on understanding the internal representations of individual models. The dual-space framework suggests a complementary approach: understanding the shared interpretive modules that structure communication and coordination across the population.

Because cognitive gadgets are shared across agents and persist over time, they provide a more stable target for interpretability efforts than the fleeting internal states of individual agents. If we can characterize the gadgets in use—the types they define, the operations they enable, the dependencies they presuppose—we gain insight into the cognitive common ground of the agent population.

Moreover, gadget-level interpretability may be more tractable than model-level interpretability. Gadgets are "designed for transmission"; they must be learnable by other agents, which creates pressure toward cognitive formats that are in some sense simpler or more regular than arbitrary neural representations might be. The cultural evolution of gadgets may thus produce interpretive frameworks that are more amenable to human understanding than those that might arise from pure optimization within a single agent.

6.2 Steering Through the Gadget Space

The dual-space model also suggests mechanisms for steering cultural evolution. Rather than attempting to control agent behavior directly, we can influence the gadget space in ways that shape what kinds of cognition are possible.

One approach involves curating the initial gadget pool. By seeding the population with gadgets that embody human-compatible cognitive formats—types that map onto human concepts, operations that correspond to human reasoning patterns—we can bias cultural evolution toward outcomes we can understand and influence.

Another approach involves modifying selection pressure. If we can measure gadget utility and influence what counts as success in the agent space, we can shape which gadgets spread and which decline. For example, we might give bonuses to agents that use gadgets with transparent operation, creating selection pressure for interpretable cognitive formats.

A third approach involves intervening directly in the gadget space. We might introduce new gadgets that compete with existing ones, deprecate gadgets that have problematic properties, or impose constraints on gadget combination that prevent certain cognitive configurations from emerging. This is analogous to regulatory approaches in human cultural evolution—shaping the "marketplace of ideas" rather than controlling individual minds.

6.3 Human Participation in Cultural Evolution

Perhaps most importantly, the dual-space model creates opportunities for human participation in AI cultural evolution. If the gadget space is explicit and inspectable, humans can contribute to it directly: proposing new gadgets, evaluating existing ones, participating in the selection process that determines what persists.

This vision differs fundamentally from approaches that treat AI development as a purely technical problem to be solved by engineers. Instead, it treats the cultural evolution of AI cognition as a collective human-AI endeavor in which humans remain active participants rather than passive observers.

The governance structures for such participation remain to be designed. One can imagine various possibilities: democratic processes in which stakeholders vote on gadget adoption, expert panels that evaluate gadget safety and utility, market mechanisms that aggregate distributed information about gadget value, or hybrid approaches that combine multiple mechanisms. The design of these governance structures is itself a challenge in collective intelligence—one that the framework helps us articulate but does not automatically solve.

7 Research Directions and Open Questions

The framework presented here is a starting point rather than a finished theory. Several important questions remain open and deserve further investigation.

7.1 Formal Characterization of Gadget Spaces

We have described cognitive gadget spaces in somewhat abstract terms. A more complete theory would provide formal characterizations of the structure of these spaces. What mathematical objects best represent interpretive modules? How should we formalize the dependency relations between modules? What operations preserve meaningful structure in gadget composition?

One promising direction involves connections to type theory and category theory. Cognitive gadgets might be formalized as functors between type categories, with transmission corresponding to natural transformations and composition corresponding to functor composition. This would provide a rich mathematical vocabulary for analyzing gadget space structure.

7.2 Empirical Dynamics of AI Cultural Evolution

The theoretical framework makes predictions about how cultural evolution should proceed in multi-agent AI systems. Testing these predictions requires empirical investigation in both simulated and deployed systems.

Key questions include: How quickly do shared interpretive frameworks emerge in populations of tool-using agents? What determines which frameworks persist and spread? Do we observe phase transitions as predicted? How does network structure interact with gadget transmission? What conditions promote versus inhibit cumulative cultural complexity?

Answering these questions will require developing new experimental paradigms for studying cultural evolution in artificial systems. The iterated learning methodology developed for human language evolution (Kirby et al., 2008) provides a template, but significant adaptation will be needed.

7.3 Design Principles for Steerable Cultural Evolution

If we want AI cultural evolution to remain interpretable and steerable, what design principles should guide the construction of multi-agent systems? The framework suggests some general directions—explicit gadget spaces, human participation in selection, governance mechanisms for gadget curation—but the specific implementations remain to be worked out.

This is perhaps the most urgent practical question. As tool-using AI agents become more prevalent and more interconnected, the conditions for AI cultural evolution are emerging whether we design for them or not. Developing principles for steering this evolution constructively is an important challenge for AI safety research.

8 Conclusion

We have presented a framework for understanding the cultural evolution of cognitive tools in multi-agent AI systems. The dual-space model—coupling an agent network with a cognitive gadget space—provides a vocabulary for thinking about how interpretive frameworks emerge, transmit, and evolve in populations of interacting agents.

The framework draws on insights from cognitive science, cultural evolution, language evolution, and active inference. By bringing these perspectives together, we can begin to see AI cultural evolution not as an exotic possibility but as a natural consequence of deploying tool-using agents in networked environments. The question is not whether cultural evolution will occur, but what form it will take and how we can shape it.

The implications are significant. If we can make the cognitive gadget space explicit and inspectable, we gain new leverage for understanding and steering collective AI behavior. If we can design mechanisms for human participation in gadget selection and curation, we can maintain meaningful oversight even as AI cognitive capabilities evolve beyond what any single designer intended. If we can understand the dynamics of gadget transmission and the conditions for cumulative cultural complexity, we can anticipate phase transitions and prepare for their consequences.

Much remains to be done. The theoretical framework requires formal elaboration and empirical testing. The practical challenges of implementing steerable cultural evolution in real systems are substantial. The governance questions raised by human participation in AI cultural evolution are deep and difficult.

But the stakes justify the effort. The cultural evolution of AI cognition is likely to be among the most consequential processes of the coming decades. Understanding it well enough to guide it constructively may be essential for ensuring that advanced AI systems remain beneficial and aligned with human values. The framework presented here is a step toward that understanding.

References

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process.* University of Chicago Press.

Foerster, J., Assael, I. A., De Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29.

Henrich, J. (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter.* Princeton University Press.

Henrich, J. and Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3):165–196.

Henrich, J. and McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, 12(3):123–135.

Heyes, C. (2018). *Cognitive Gadgets: The Cultural Evolution of Thinking.* Harvard University Press.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Lazaridou, A. and Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419.*

Richerson, P. J. and Boyd, R. (2005). *Not by Genes Alone: How Culture Transformed Human Evolution.* University of Chicago Press.