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 experience[13]—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 planning[20] and cognitive architectures such as ACT-R[21] and SOAR[22] 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).

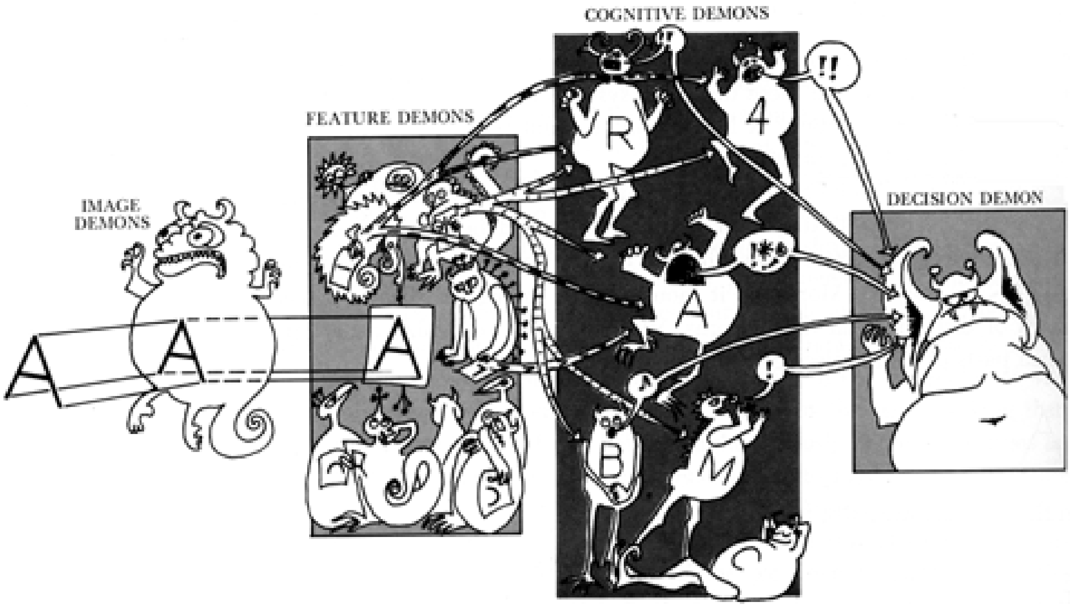

Figure 1 An illustration of Oliver Selfridge's 1959 pandemonium architecture model, drawn by Leanne Hinton. Source: Lindsay and Norman, Human Information Processing.

[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).

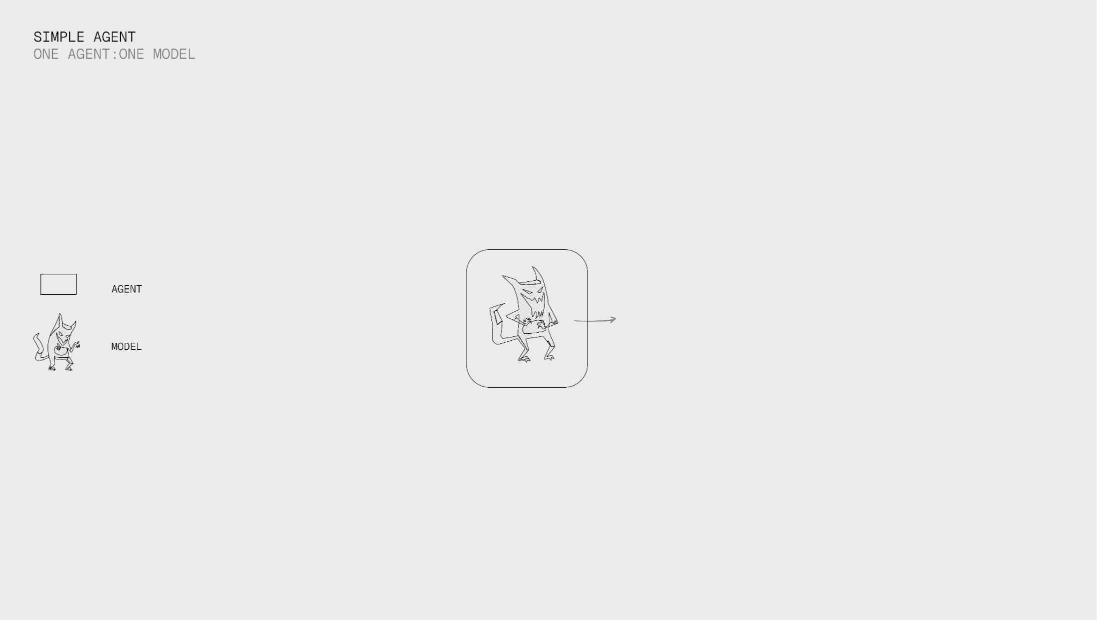

Figure 2 A simple agent composed of one model.

In contrast, in pandemonium architecture, multiple models, or *daemons*, coexist within a single agent, each processing information or generating responses independently (Figure 3).

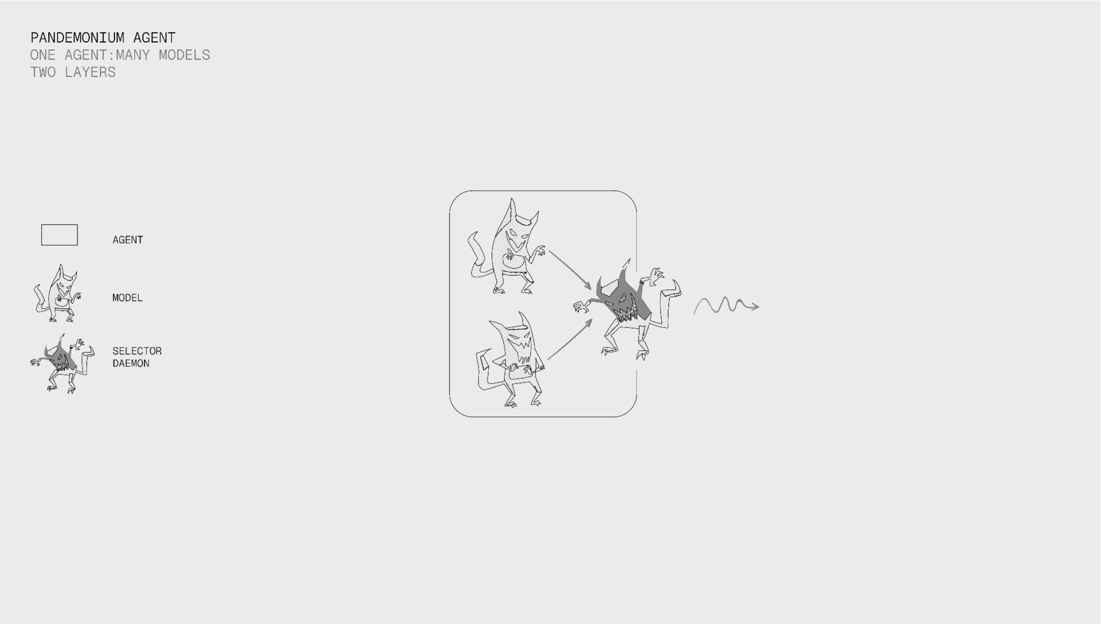

Figure 3 A pandemonium agent composed of two models and a selector daemon.

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

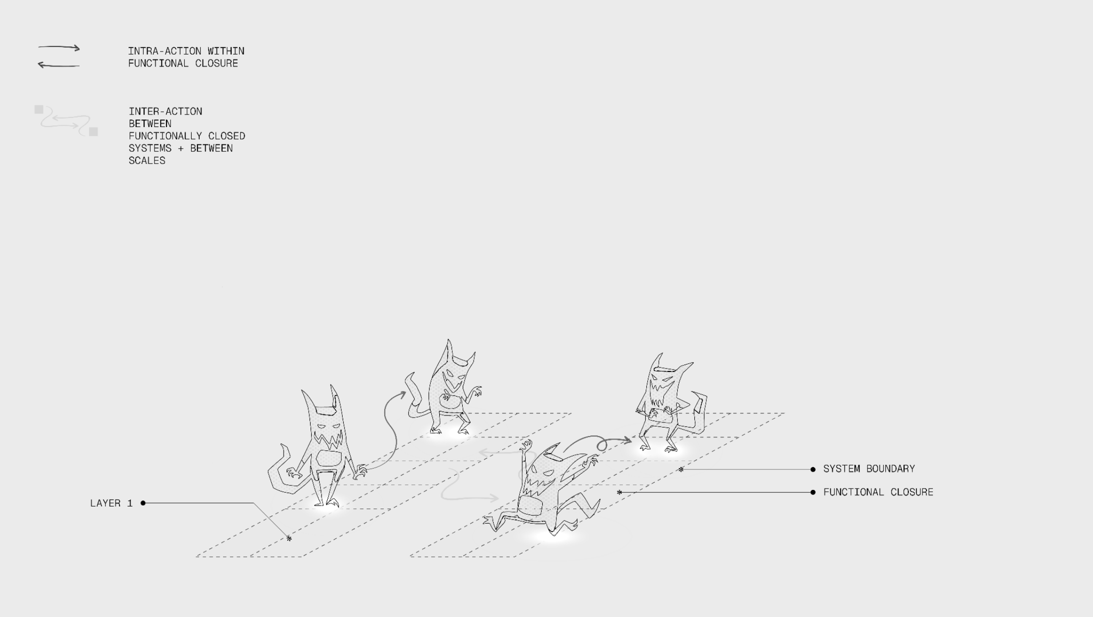

Figure 4 Inter-action and intra-action for two distinct functionally closed layers within an agent.

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*).

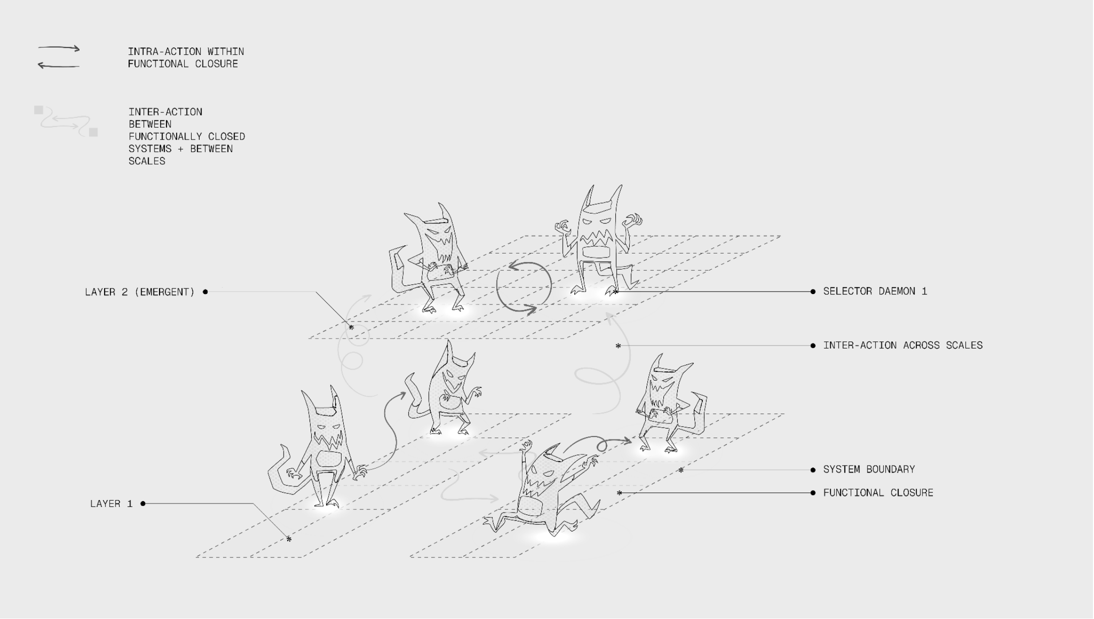
INTRA-ACTION WITHIN FUNCTIONAL CLOSURE
INTER-ACTION BETWEEN FUNCTIONALLY CLOSED SYSTEMS + BETWEEN SCALES
LAYER 2 (EMERGENT)
SELECTOR DAEMON 1
INTER-ACTION ACROSS SCALES
SYSTEM BOUNDARY
FUNCTIONAL CLOSURE
LAYER 1

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).

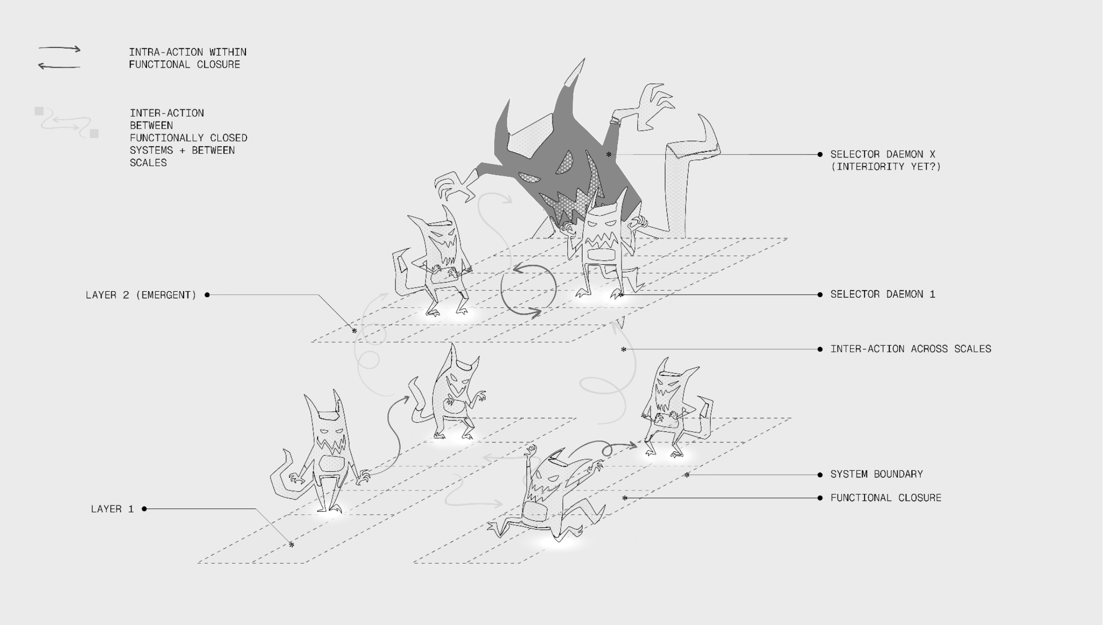
INTRA-ACTION WITHIN FUNCTIONAL CLOSURE
INTER-ACTION BETWEEN FUNCTIONALLY CLOSED SYSTEMS + BETWEEN SCALES
SELECTOR DAEMON X (INTERIORITY YET?)
LAYER 2 (EMERGENT)
SELECTOR DAEMON 1
INTER-ACTION ACROSS SCALES
SYSTEM BOUNDARY
FUNCTIONAL CLOSURE
LAYER 1

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.

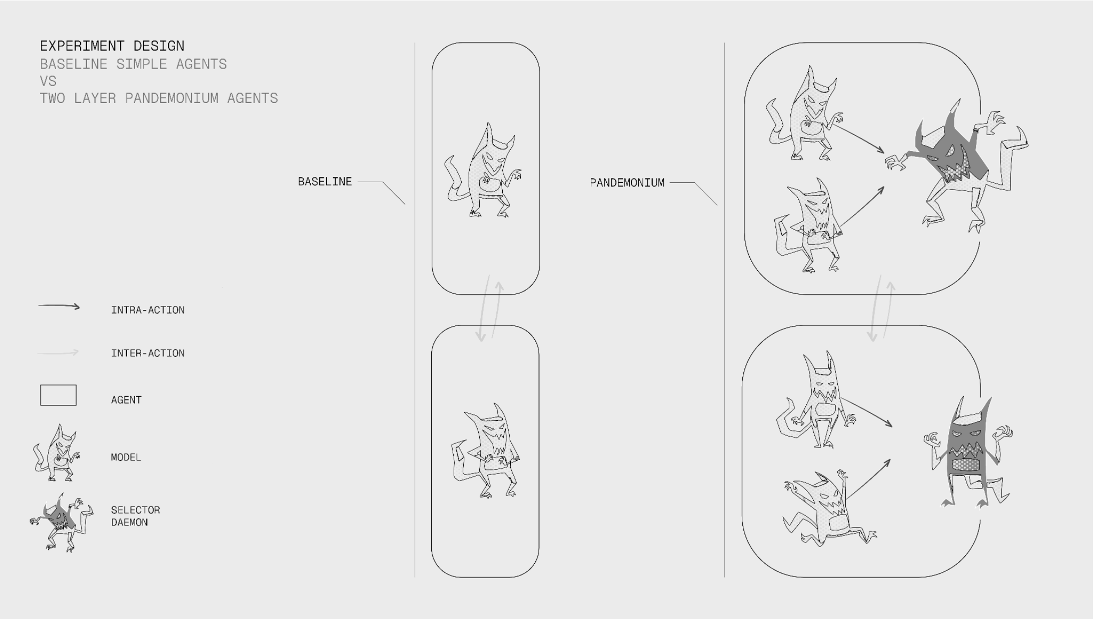

Figure 7 Comparison of the baseline scenario with a pandemonium scenario with one layer of functional closure, for N=2.

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.

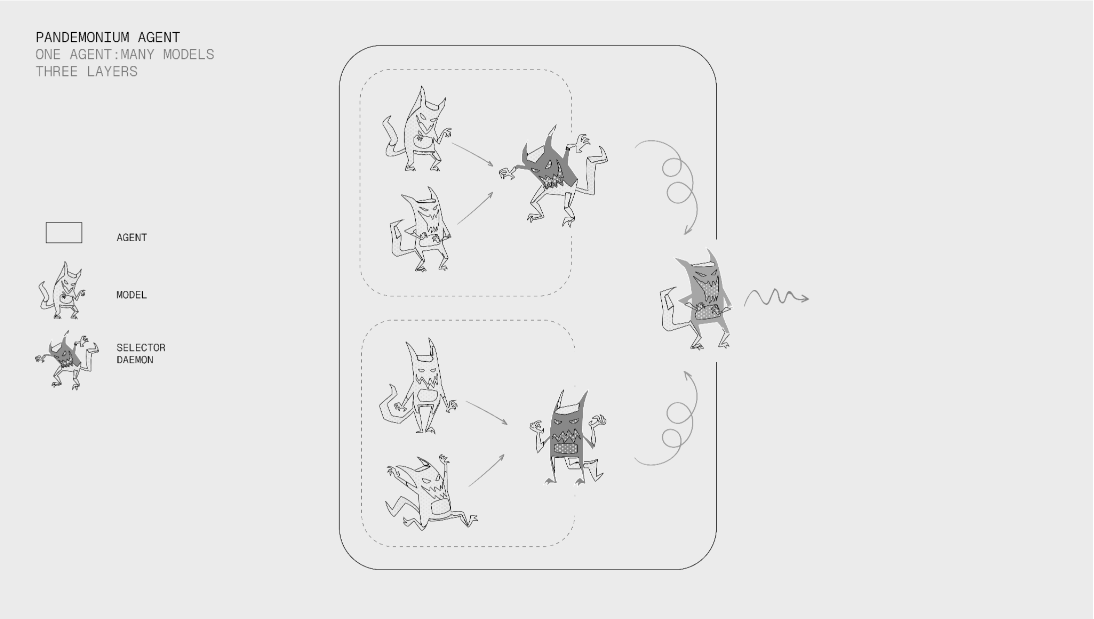

Figure 8 A single agent containing a pandemonium architecture with three layers.

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).

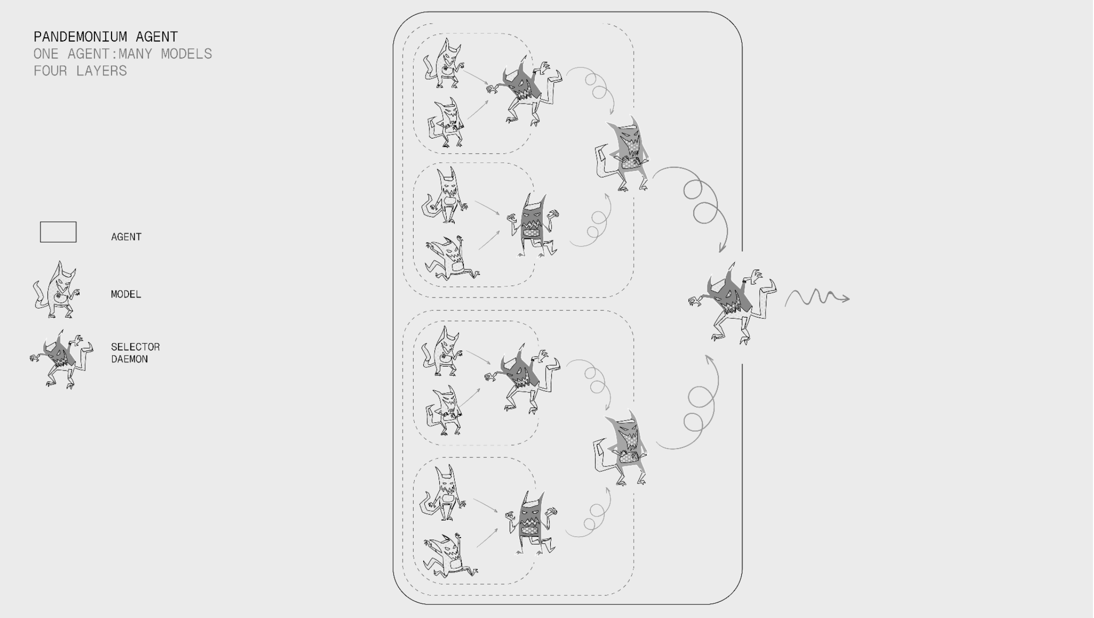

Figure 9 A single agent containing a pandemonium architecture with four layers

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.

Warpefelt, 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.