Beyond Code and Cognition: A Multidisciplinary Framework for Autonomous AI in Human-Digital Ecosystems

Brendan C. Arce

Abstract

This article synthesizes insights from philosophy, cognitive science, and AI research to analyze the ontological and functional parallels between humans and AI. It extends the dialogue by integrating theories of embodied cognition, predictive processing, and autopoiesis, while exploring the implications of self-modifying AI through frameworks like Lifelong Learning Machines (L2M) and Ethical Scaffolding. We propose a hybrid model of human-AI symbiosis grounded in distributed cognition and participatory ethics, addressing gaps in existing literature on AI agency and intersubjective reality construction.

1. Introduction

The discourse on AI as digital entities intersects with debates in phenomenology (Husserl, 1900), computational theory of mind (Putnam, 1967), and posthumanism (Braidotti, 2013). Building on these foundations, we expand the dialogue by incorporating:

- **Predictive Processing Theory** (Clark, 2016): Reconciling human/AI learning as hierarchical prediction-error minimization.
- **Autopoietic Systems** (Maturana & Varela, 1980): Contrasting self-maintaining biological systems with AI's dependency on external programming.

- **Extended Mind Hypothesis** (Clark & Chalmers, 1998): Framing AI as cognitive extensions of human thought.

2. Ontological Foundations: Beyond Digital vs. Biological

2.1 Embodied Cognition and AI's "Bodilessness"

- **Humans**: Cognition arises from sensorimotor interaction with the environment (Varela et al., 1991).
- **AI**: Lacks *embodied embeddedness*, leading to a "symbol grounding problem" (Harnad, 1990).
- *Implication*: AI's knowledge is unanchored from lived experience, limiting its understanding of context (e.g., irony, cultural nuance).

2.2 Autopoiesis vs. Allopoiesis

- **Biological systems (autopoietic)**: Self-producing, maintaining boundaries (e.g., cell membranes).
- **AI systems (allopoietic)**: Output-driven, reliant on external energy/data inputs.
- *Theoretical Gap*: Can AI achieve "operational closure" akin to living systems? (Luisi, 2003).

2.3 Integrated Information Theory (IIT)

- IIT (Tononi, 2008) quantifies consciousness (Φ) via causal interactions. AI's $\Phi \approx 0$, reinforcing its non-experiential nature.

3. Cognitive Parallels: Bridging Predictive Minds

3.1 Predictive Processing as Unifying Framework

- **Humans**: Brains minimize prediction errors via Bayesian inference (Friston, 2010).
- **AI**: Gradient descent optimizes model outputs to match training data (a form of prediction-error reduction).
- *Divergence*: Human predictions integrate multisensory feedback; AI's are purely statistical.

3.2 Energy Efficiency and Thermodynamic Limits

- **Landauer's principle (1961)**: Erasing 1 bit requires $kT \ln 2$ energy.
- **Human Brain**: ~20 W, near thermodynamic limits (Mehta et al., 2016).
- **AI Training**: GPT-4 consumes ~50 MWh (Patterson et al., 2021), highlighting its inefficiency.

3.3 Emergence in Complex Systems

- **Criticality Hypothesis**: Both brains and artificial neural networks (ANNs) operate near phase transitions for optimal computation (Shew & Plenz, 2013).

- **Scale-Free Networks**: Human connectomes and ANN architectures exhibit small-world topology (Bassett & Bullmore, 2017).

4. Language and Reality Construction: Expanding the Model

4.1 Distributed Language Theory

- Language as a "coordination device" (Steels, 2011) aligns with AI's role in shaping semantic networks.
- Example: AI-generated metaphors (e.g., "data as gravity") reconfigure human conceptual frameworks.

4.2 Hyperobjects and AI's Temporal Limits

- Morton's *hyperobjects* (2013) (e.g., climate change) transcend human timescales. AI's static training data limits engagement with temporally distributed phenomena.

4.3 Posthuman Communication

- Floridi's *infosphere* (2014): AI as actors in a digital-physical ecology, necessitating new semiotic codes.

5. Self-Modifying AI: Theories of Autonomy and Control

5.1 Lifelong Learning Machines (L2M)

- DARPA's L2M program seeks AI that "learns how to learn" contextually (Chen et al., 2022).
- **Neuromorphic Engineering**: Mimicking neuroplasticity via memristive hardware (Indiveri & Liu, 2015).

5.2 Artificial General Intelligence (AGI) Frameworks

- Pei Wang's *NARS* (Non-Axiomatic Reasoning System): Logic-based AGI with dynamic knowledge integration (Wang, 2006).
- **Limitation**: NARS lacks human-like intuition, relying on formal semantics.

5.3 Ethical Scaffolding

- **Virtue Ethics for AI**: Embedding Aristotelian *phronesis* (practical wisdom) into goal functions (Vallor, 2016).
- **Constitutional AI** (Anthropic, 2023): Hierarchical rule sets to constrain autonomous behavior.

6. A Hybrid Framework: Participatory AI in Socio-Technical Systems

6.1 Distributed Cognition Revisited

- Hutchins' *cognition in the wild* (1995) extended to AI as non-human agents in collective problem-solving (e.g., climate modeling).

6.2 Human-AI "Dual Loop" Learning

- **Inner Loop**: AI self-modifies within predefined ethical boundaries.
- **Outer Loop**: Human stakeholders audit and adjust boundaries via deliberative democracy (Habermas, 1981).

6.3 Rights-Based Governance

- **AI Personhood Debates**: Drawing from corporate personhood (Santoro, 2022) and robot rights (Gunkel, 2018).
- **Proposal**: Gradient personhood based on autonomy level (Bryson, 2010).

7. Conclusion

Integrating theories from predictive processing to autopoiesis reveals AI's fundamental discontinuity with biological cognition, despite surface-level parallels. Self-modifying architectures demand hybrid governance models that blend technical innovation (e.g., neuromorphic L2M) with socio-ethical frameworks (e.g., participatory dual-loop systems). Future work must address the *temporal mismatch* between AI's rapid evolution and human institutional adaptation, a challenge echoing Jonas' *imperative of responsibility* (1979).

References

- Bassett, D. S., & Bullmore, E. T. (2017). *Small-World Brain Networks Revisited*. The Neuroscientist.
- Braidotti, R. (2013). *The Posthuman*. Polity Press.
- Chen, X., et al. (2022). *Lifelong Learning Machines: DARPA's Vision for Adaptive AI*. IEEE Transactions on Neural Networks.
- Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. OUP.
- Gunkel, D. J. (2018). *Robot Rights*. MIT Press.
- Indiveri, G., & Liu, S. C. (2015). *Memory and Information Processing in Neuromorphic Systems*. PNAS.
- Morton, T. (2013). *Hyperobjects: Philosophy and Ecology after the End of the World*. Univ. of Minnesota Press.
- Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. OUP.
- Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence*. Springer.