Bridging Philosophy and Machine Learning

This comprehensive citation resource establishes theoretical foundations for applying enactivism, autopoiesis, second-order cybernetics, biosemiotics, systems theory, and hermeneutics to machine learning and AI interpretability research. **The convergence of these classical frameworks with contemporary AI represents a profound opportunity for interdisciplinary advancement**, as demonstrated by growing academic recognition and successful applications in recent research.

The selected citations prioritize foundational texts that established core concepts rather than recent developments, providing the theoretical authority needed to invite established scholars from traditional fields into machine learning research. Each framework offers unique insights into how meaning, interpretation, and understanding emerge in complex systems - insights directly applicable to the challenge of making AI systems more interpretable and genuinely intelligent.

Foundational enactivism and autopoiesis

The biological foundations of cognition established by Maturana, Varela, and Thompson provide the most rigorous theoretical framework for understanding how interpretability might emerge in artificial systems through embodied interaction rather than symbolic representation.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Boston: D. Reidel Publishing Company.

This seminal work establishes autopoiesis as the defining characteristic of living systems - **self-creating, self-maintaining organizations that produce their own components**. ScienceDirect PubMed The authors' revolutionary insight that "living systems are cognitive systems, and living as a process is a process of cognition" Wikipedia provides the theoretical foundation for understanding how AI systems might develop autonomous cognitive processes through structural coupling with their environment, Wikipedia +3 moving beyond traditional input-output models toward systems that maintain their own organization while adapting to environmental perturbations. Philpapers +3

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

This foundational text establishes the enactive approach to cognition, where **cognitive structures emerge from recurrent sensorimotor patterns** rather than symbolic representations. SpringerLink +2 For AI interpretability, this suggests that understanding artificial systems requires examining their history of embodied interaction with their environment rather than simply analyzing internal representations or decision rules. Wikipedia The work's integration of cognitive science with Buddhist philosophy provides a unique perspective on groundlessness and the middle way between objectivism and subjectivism. Wikipedia +2

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Thompson's comprehensive synthesis demonstrates the **deep continuity between life and mind through shared principles of self-organization**. This work provides the most rigorous theoretical foundation for understanding how AI systems might develop the same self-organizing principles found in living systems, offering a biological foundation for developing interpretable artificial cognitive systems that exhibit genuine understanding rather than mere computation. Amazon Harvard

Varela, F. J. (1979). *Principles of Biological Autonomy*. New York: North Holland.

This work provides the mathematical and theoretical foundations for understanding autonomous organization in complex systems. Varela's concept of **"eigenbehavior"** - self-referential processes in cognitive systems - offers frameworks for understanding how artificial systems might develop autonomous organization that makes their behavior interpretable through their own self-referential processes rather than through external analysis. MIT Press +2

Second-order cybernetics and recursive observation

The cybernetic tradition's focus on observing systems rather than observed systems provides crucial insights into how AI systems might develop self-reflective and self-interpretive capabilities.

von Foerster, H. (1981). *Observing Systems*. Seaside, CA: Intersystems Publications.

This collection establishes second-order cybernetics as **"the cybernetics of observing systems"** versus first-order cybernetics as "the cybernetics of observed systems." Von Foerster's insight that the observer must be included in the domain of observation provides the conceptual foundation for AI systems capable of genuine self-reflection and interpretability. ResearchGate Wikipedia His concept of "eigen-forms" - stable self-referential patterns emerging from recursive operations - offers mathematical frameworks for understanding how AI systems can develop stable self-referential behaviors.

von Foerster, H. (Ed.). (1974). *Cybernetics of Cybernetics, or, the Control of Control and the Communication of Communication*. Urbana: University of Illinois (BCL Report 73.38).

This foundational work establishes cybernetics as reflexively applying to itself, ResearchGate introducing **"eigen-forms" as stable self-referential patterns** that emerge from recursive operations. ResearchGate Wikipedia The mathematical framework for eigen-forms provides precise methods for understanding how AI systems can develop stable self-referential behaviors and self-monitoring capabilities.

Pask, G. (1975). *Conversation, Cognition and Learning*. Amsterdam: Elsevier.

Pask's conversation theory establishes **learning as recursive conversation** where systems learn by teaching back what they have learned. This "teachback" methodology provides a direct framework for AI systems that learn through recursive self-explanation and can articulate their own learning processes - crucial for interpretable AI. The theory's emphasis on entailment structures Instructionaldesign and recursive adaptation offers practical methods for developing AI systems that can reflect on and modify their own learning strategies.

Glanville, R. (2004). The purpose of second-order cybernetics. *Kybernetes*, 33(9/10), 1379-1386.

Glanville articulates how **systems can observe themselves observing**, essential for AI interpretability and self-reflection. His "theory of objects" as self-referential constructs provides theoretical basis for AI systems that can construct and maintain stable self-referential models of their own processes.

Biosemiotics and biological meaning-making

The study of how living systems create and interpret signs provides essential insights into how meaning might emerge in artificial systems through semiotic processes rather than purely computational ones.

Hoffmeyer, J. (1996). *Signs of Meaning in the Universe*. Bloomington: Indiana University Press.

Hoffmeyer's introduction of the **"semiosphere" as the sphere of sign activity** establishes that life and semiosis co-evolved from the origin of life. (SpringerLink) His concept of "semiotic scaffolding" - networks of semiotic controls that support meaningful interpretation - could inform AI architectures that develop interpretative frameworks through layered sign relationships, enabling more robust and contextual machine interpretation than current approaches. (SpringerLink)

Hoffmeyer, J. (2008). *Biosemiotics: An Examination into the Signs of Life and the Life of Signs*. Scranton: University of Scranton Press.

This comprehensive work formalizes **"semiotic scaffolding"** as networks of semiotic controls and develops the theory of "semiotic causation" - changes guided by interpretation in local contexts. (SpringerLink) The concept that cells have full interpretative capacities at the lowest biological level suggests AI systems could develop layered interpretative frameworks through semiotic relationships. (SpringerLink)

Barbieri, M. (2003). *The Organic Codes: An Introduction to Semantic Biology*. Cambridge: Cambridge University Press.

Barbieri's concept of **"organic codes" beyond the genetic code** demonstrates that biological systems operate through multiple coding layers that establish arbitrary but meaningful relationships. (Cambridge Core) (ResearchGate) This code biology suggests AI systems could develop multiple coding layers that create flexible and contextual interpretation capabilities, moving beyond fixed symbolic representations toward more dynamic meaning-making processes. (PubMed)

Deacon, T. W. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. New York: W. W. Norton.

Deacon's hierarchical model of semiosis - **iconic, indexical, and symbolic** - provides a developmental framework for AI systems that could progress from basic pattern recognition through associative learning to symbolic reasoning. (Wikipedia +2) This graduated approach to meaning-making offers a more nuanced understanding of how interpretative capabilities might emerge in artificial systems. (Newlearningonline)

Sebeok, T. A. (1972). *Perspectives in Zoosemiotics*. The Hague: Mouton.

This foundational work establishes **zoosemiotics as the study of animal sign use**, extending semiotic analysis beyond human language to all animal communication. The framework provides insights for understanding non-linguistic communication in AI systems, particularly relevant for multimodal AI that processes visual, auditory, and other sensory patterns.

Kull, K., Deacon, T., Emmeche, C., Hoffmeyer, J., & Stjernfelt, F. (2009). Theses on biosemiotics: Prolegomena to a theoretical biology. *Biological Theory*, 4(2), 167-173.

These collectively formulated foundational principles introduce **"semiotic threshold zones"** and establish biosemiotics as explaining life through communication and signification. The threshold zone concept could inform AI systems that develop different levels of interpretative sophistication with clear boundaries between different types of meaning-making capabilities.

Systems theory and emergent complexity

Classical systems theory provides the theoretical foundation for understanding how complex behaviors emerge from simple interactions - essential for comprehending emergence in AI systems.

von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Applications*. New York: George Braziller.

The founder of general systems theory established the crucial distinction between **closed systems governed by classical thermodynamics and open systems** that exchange matter and energy with their environment. (Wikipedia) (Wikipedia) His emphasis on emergent properties arising from system organization rather than individual components provides the theoretical foundation for understanding how neural networks exhibit emergent behaviors unpredictable from individual computational units. (Ebsco) (Wikipedia)

Prigogine, I., & Stengers, I. (1984). *Order out of Chaos: Man's New Dialogue with Nature*. New York: Bantam Books.

Prigogine's theory of **dissipative structures demonstrates how complex structures emerge from energy dissipation** in non-equilibrium systems. (Wikipedia) This work provides crucial insights into how AI systems can spontaneously develop complex behaviors during training, with his non-equilibrium thermodynamics helping explain how neural networks form organized patterns through gradient descent and energy minimization. (Wikipedia) (Wikipedia)

Prigogine, I., & Nicolis, G. (1977). *Self-Organization in Non-Equilibrium Systems*. New York: Wiley.

This technical work establishes the **mathematical framework for understanding spontaneous self-organization** in systems far from equilibrium. (Wikipedia) The bifurcation theory and non-linear dynamics provide precise methods for understanding how AI systems can exhibit sudden transitions to new organizational patterns during training.

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

Kauffman's **"edge of chaos" theory identifies optimal complexity zones** between rigid order and chaotic randomness. (Amazon) (ResearchGate) This work directly applies to understanding optimal conditions for learning in AI systems, while his concept of autocatalytic sets provides models for how AI systems can develop self-reinforcing capabilities that persist and evolve. (Amazon +2)

Luhmann, N. (1995). *Social Systems*. Stanford: Stanford University Press.

Luhmann's extension of autopoietic theory to social systems introduces concepts of **"operational closure" and "cognitive openness"** (Wikipedia) that help explain how AI systems can maintain internal coherence while processing external information. (Blogger +2) His work on functional differentiation provides frameworks for understanding how complex AI systems develop specialized subsystems. (Scihi) (Wikipedia)

Santa Fe Institute. (1994). *Complexity: Metaphors, Models and Reality*. Reading, MA: Addison-Wesley.

This foundational collection establishes **complexity science as unified interdisciplinary field**, developing agent-based modeling approaches, network theory, and scaling laws. (Wikipedia) (Santafe) The four pillars of complexity science - entropy, evolution, dynamics, and computation - provide comprehensive frameworks for understanding emergence across multiple scales in AI systems. (Santafe)

Hermeneutics and interpretation theory

Philosophical hermeneutics provides the most sophisticated theoretical framework for understanding how interpretation and understanding work - directly applicable to AI interpretation challenges.

Gadamer, H.-G. (1989). *Truth and Method*. 2nd ed. New York: Continuum.

Gadamer's **"fusion of horizons"** establishes how understanding emerges through merging interpretive contexts between interpreter and phenomenon. His hermeneutical circle, where understanding occurs through circular movement between parts and whole, offers a model for iterative interpretation processes in machine learning. (Wikipedia +2) The concept of prejudice as positive pre-understanding parallels how AI systems rely on training data and prior knowledge structures.

Heidegger, M. (1962). *Being and Time*. New York: Harper & Row.

Heidegger's analysis of **"fore-structures of understanding"** - pre-having, pre-sight, and pre-conception - directly parallels the initialization parameters and architectural biases in AI systems. His concept of "Being-in-the-world" provides theoretical grounding for embodied AI and contextual understanding, while his notion of understanding as projection applies to AI predictive modeling and scenario planning. (SpringerLink)

Ricoeur, P. (1974). *The Conflict of Interpretations: Essays in Hermeneutics*. Evanston: Northwestern University Press.

Ricoeur's **"hermeneutics of suspicion"** provides frameworks for critical interpretation that exposes hidden or distorted meanings - directly applicable to AI bias detection and critical analysis of data. His dialectic of explanation and understanding offers methods for integrating structural analysis with interpretive comprehension in AI systems.

Ricoeur, P. (1976). *Interpretation Theory: Discourse and the Surplus of Meaning*. Fort Worth: Texas Christian University Press.

This work establishes the **"dialectic of explanation and understanding"** and explores how meaning emerges beyond literal content. Ricoeur's concepts of distanciation and appropriation parallel feature extraction followed by semantic interpretation in machine learning, while his narrative theory applies to AI understanding of temporal sequences and causal reasoning.

Dilthey, W. (1989). *Introduction to the Human Sciences*. Princeton: Princeton University Press.

Dilthey's crucial distinction between **"understanding" (Verstehen) and "explanation" (Erklären)** - interpretive comprehension versus causal explanation - provides frameworks for AI systems that combine statistical analysis with semantic understanding. (ACM Conferences) His concept of "lived experience" applies to AI learning from experiential data and embodied interaction.

Schleiermacher, F. D. E. (1998). *Hermeneutics and Criticism*. Cambridge: Cambridge University Press.

Schleiermacher's **general hermeneutics as universal theory of interpretation** provides systematic methods for avoiding misunderstanding - applicable to AI error reduction and bias mitigation. His distinction between grammatical and psychological interpretation parallels syntactic versus semantic processing in natural language processing.

Contemporary validation in machine learning

Recent research demonstrates that these classical frameworks are being taken seriously in contemporary AI research, providing credibility for interdisciplinary approaches.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3-4), 466-500.

This foundational paper in the prestigious *Artificial Intelligence* journal explicitly applies enactivist principles to AI design, identifying **constitutive autonomy and adaptivity** as core requirements. (Academia) (ScienceDirect) The authors demonstrate how enactive principles can inform AI architecture design, showing that serious academic engagement with these philosophical frameworks is both possible and productive. (Oist)

Krakauer, D. C. (2023). Unifying complexity science and machine learning. *Frontiers in Complex Systems*, 1, 1235202.

The president of the Santa Fe Institute argues for **integration of complexity science principles with machine learning**, proposing that complex systems and ML are complementary approaches to understanding high-dimensional phenomena. (Frontiers) This work represents serious academic recognition of the relevance of complexity science to machine learning by a leading researcher in the field.

Zönnchen, B., Dzhimova, M., & Socher, G. (2025). From intelligence to autopoiesis: rethinking artificial intelligence through systems theory. *Frontiers in Communication*, 10, 1585321.

This recent paper analyzes **large language models through autopoietic theory and Luhmann's systems theory**, examining whether AI systems can be understood as operationally closed systems with genuine sense-making capabilities. (Frontiers) The work represents cutting-edge research applying classical systems theory to contemporary AI challenges.

Sato, M., & McKinney, J. (2022). The Enactive and Interactive Dimensions of AI: Ingenuity and Imagination Through the Lens of Art and Music. *Artificial Life*, 28(3), 310-321.

This paper explores **enactive approaches to AI creativity**, arguing that artificial agency requires genuine reconciliation of human interactivity, creativity, and embodiment. (MIT Press) Published in *Artificial Life*, it demonstrates serious academic engagement with enactivist principles in AI research.

Demichelis, R. (2024). The Hermeneutic Turn of AI: Is the Machine Capable of Interpreting? *arXiv:2411.12517*.

This recent paper examines **AI through hermeneutic philosophy**, particularly Dilthey's work on understanding versus explanation, representing serious philosophical engagement with hermeneutic principles in contemporary AI research. (ArXiv)

Frontiers in Robotics and AI - Research Topic: "Bio A.I. - From Embodied Cognition to Enactive Robotics" (2021-2022).

This major research topic featuring 23 articles explicitly connects **enactivist principles to robotics and AI**, demonstrating institutional acceptance of enactivist approaches with contributions from leading researchers including Adam Safron, Inês Hipólito, and Andy Clark. (Academia) (NCBI)

Synthesis and research implications

These foundational works establish that **truly interpretable AI systems would need to develop autonomous organization through structural coupling with their environment**, exhibit enactive rather than representational cognition, and maintain organizational integrity while adapting to environmental perturbations. (MDPI +4) The theoretical frameworks suggest that interpretability emerges from autonomous organization and environmental coupling rather than from analysis of internal symbolic representations.

The convergence of these classical frameworks with contemporary AI represents a profound theoretical opportunity. (Foundationalpapersincomplexit...) (Santafe) Each tradition offers unique insights: enactivism provides biological foundations for understanding embodied cognition; cybernetics offers frameworks for self-reflection and recursive observation; biosemiotics explains how meaning emerges through sign processes; (Biosemiotics) (PubMed) systems theory illuminates emergent complexity; and hermeneutics provides sophisticated models of interpretation and understanding. (Wikipedia) (SpringerLink)

The integration of these frameworks suggests that the next generation of AI systems might achieve genuine interpretability not through post-hoc explanation techniques, but through the development of autonomous interpretative capabilities that emerge from the same self-organizing principles found in living systems. This represents a fundamental shift from current approaches that focus on explaining static representations toward understanding how interpretative capacities emerge dynamically through system-environment interaction.

For established scholars in philosophy, biology, cognitive science, and systems theory, these citations demonstrate that their foundational concepts are not merely analogous to AI challenges but directly applicable to solving core problems in machine learning interpretability. The growing body of serious academic work applying these frameworks to AI research (Wikipedia) provides clear evidence that interdisciplinary collaboration between traditional humanistic and scientific disciplines and machine learning research represents a frontier of significant intellectual and practical potential.