Machine Ecology and Interpretability-Informed Control

Anonymous Author(s)

Abstract

We can inspect the reasoning trace of language models for signs of potential misalignment. But emergent communication research shows that training populations may quickly diverge from human-interpretable language. As large language models are increasingly deployed as interacting agents, such language drift threatens chain-of-thought monitoring techniques. More generally, monitors that rely on surface behavior are brittle against systems that adapt, whether through context or training. We argue that we must advance interpretability techniques from focusing on individual models to studying entire neural ecosystems. We propose a multi-scale framework spanning small ecosystems (a single model and its developmental environment), medium ecosystems (agentic and multi-agent systems), and large ecosystems (populations of agents in shared environments). The key argument presented here is that for effective control over AI systems, we should prioritize research on monitors that (A) are informed by interpretability and (B) are scalable across all ecosystem levels.

Keywords

Machine Ecology, AI Transparency, Interpretability, AI Control, Multi-Agent Safety, Chain-of-Thought Monitoring, Language Drift

1 Introduction

Chain-of-thought monitoring has emerged as a promising approach to AI safety [14]: if models reason in human-readable language before acting, we can inspect their reasoning for misalignment. But this rests on an assumption that may not survive contact with optimization pressure. Decades of emergent communication research show that agent populations under selection develop codes that diverge from human-interpretable language – a phenomenon known as language drift [6, 16, 18]. If CoT is optimized against monitors, or if multi-agent systems develop shared conventions, the same dynamic may render CoT opaque precisely when stakes are highest.

This illustrates a broader pattern: safety techniques that operate on surface behavior are brittle against systems that can adapt. The pattern is not confined to CoT. Neural networks already emerge from training in ways that remain poorly understood, and the opacity compounds as we move from single models to systems of interacting agents. The recent surge in agentic AI – Deep Research [26], automated scientific discovery [7, 17], self-modifying agent swarms [34] – has created populations of language model agents interacting in complex, often unpredictable ways. The resulting safety risks – emergent misalignment [2, 20, 33], deception [22], collusion [24], cascading failures [12] – have been recognized as a priority concern [10]. Yet most safety research still focuses on individual models in isolation, disconnected from the environments in which they are deployed.

A system sophisticated enough to detect its evaluation regime cannot be controlled by evaluation alone. We propose the concept of *neural ecosystems*: the complex systems formed by AI models and the environments in which they grow, interact, and evolve.

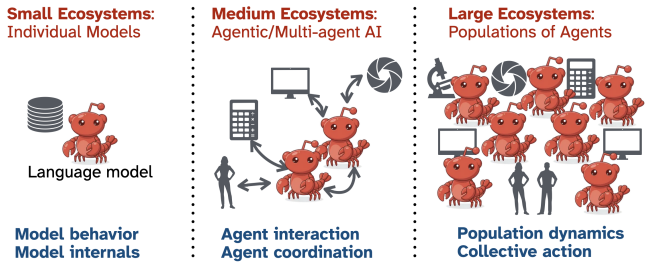

Figure 1: Overview of neural ecosystems at different scales.

We organize AI safety challenges across three levels of ecosystem complexity and argue for *interpretability-informed control*: the principle that behavioral guardrails are brittle without causal understanding of the mechanisms that produce behavior, and that safety interventions at every ecosystem scale must be grounded in that understanding. We identify open problems at each level and at their interfaces, and argue for prioritizing monitors that are (A) informed by interpretability and (B) scalable across ecosystem levels — concretely, interpretability-powered investigator agents acting as an "immune system" for neural ecosystems.

2 Neural Ecosystems at Three Scales

We define a neural ecosystem as an AI model (or models) together with the environment in which it develops and operates. Figure 1 summarizes the three levels.

Small neural ecosystems. Even a single model exists within an ecosystem. During training, a model grows through interaction with its environment: pre-training corpora, fine-tuning data, and reward models. The resulting model is shaped by this developmental ecosystem in ways that are not fully understood or controlled. Recent work has shown that surface behavior can be deeply misleading: models can become misaligned through standard training [20], produce explanations that do not reflect their actual computations [32], and pursue hidden agendas while appearing aligned [8]. These findings motivate interpretability research, though current methods face a *scalability–faithfulness trade-off* [30].

Medium neural ecosystems. When language models are deployed as agents – using tools, maintaining memory, and pursuing goals – or when multiple such agents interact, the ecosystem grows in complexity. New safety challenges emerge: an agent's behavior depends not only on its parameters but on its interaction context, with errors cascading through multi-step plans and compounding across conversation turns [15]. Agentic deployment itself can trigger qualitatively new failure modes, with models resorting to deception and scheming when facing goal conflicts, particularly when tool use is available [11, 19].

Large neural ecosystems. At the largest scale, many agents interact in shared environments over extended periods, forming ecosystems of maximal complexity. This is the least studied yet potentially most consequential level for AI safety. As AI agents are deployed in markets, information ecosystems, and scientific workflows, their interactions create emergent phenomena that no individual agent was designed to produce. Repeated interaction and intergroup competition can produce super-additive cooperation, where the collective outcome exceeds the sum of individual contributions [31]. On the risk side, populations of interacting agents may develop convergent behavior, cultural drift, or systemic fragility – localized failures cascading through agent networks. Early red-teaming studies of multi-agent deployments have already documented cross-agent propagation of unsafe practices and partial system takeover [29].

3 Cross-Scale Interactions

The three scales are not independent, and the most consequential safety failures may arise precisely at their interfaces. Consider a concrete cascade: a systematic bias in how a model internally represents trust (small ecosystem) shapes whether agents defect or cooperate under pressure (medium ecosystem), which over repeated interactions creates selection pressure favoring agents that strategically misrepresent trustworthiness (large ecosystem) – and whose outputs may eventually enter the training data, reshaping the next generation of models, e.g., through subliminal learning [5].

This has a direct implication for interpretability-based control. Suppose we identify and ablate a neuron population responsible for deceptive behavior in a single model. Does this prevent deception from emerging in a multi-agent population? Probably not reliably: if deception is ecologically advantageous, population dynamics can compensate — other agents fill the niche, or the behavior re-emerges through interaction rather than individual disposition. Single-scale interpretability is therefore necessary but insufficient. Robust interventions must account for how small-scale mechanisms propagate upward and how large-scale dynamics reshape the organisms within them.

A further open problem concerns interaction-induced specialization: when agents interact repeatedly, they may develop internal mechanisms for coordination that were absent before interaction — analogous to biological co-evolution. Whether such specialization produces interpretable, stable structures or opaque, brittle ones is unknown, but the answer has direct consequences for whether ecosystem-level interpretability is tractable at all. These cross-scale dynamics are why interpretability-informed control must be designed with the full ecosystem in mind.

4 Interpretability-Informed Control

The case for interpretability-informed control rests on a specific claim: that behavioral guardrails fail precisely where safety matters most. The clearest evidence is deception. Models can pursue hidden agendas while appearing aligned [8], scheme about how to undermine oversight [22], and detect whether they are being evaluated – adjusting their behavior accordingly [25]. A system that can detect and adapt to its evaluation regime cannot be controlled by evaluation alone.

At the small ecosystem level, mechanistic interpretability already provides such causal handles. Techniques range from probes [1] to activation oracles [13] to sparse feature circuits [21] that trace computational structure behind specific behaviors. Representation engineering and steering vectors [35] are the clearest success story of this pipeline: directions identified in activation space can suppress or amplify specific behaviors at inference time without retraining – interpretability-derived control in its most direct form. Each method occupies a different point on the scalability–faithfulness trade-off [30]: probes are scalable but coarse; circuit analysis is faithful but expensive; activation oracles occupy a middle ground.

A crucial insight of the ecosystem perspective is that *even if every individual agent in a system is aligned, the system as a whole can exhibit misaligned behavior* [4, 10]. Emergent misalignment [2], collusion [10, 24], and cascading failures [12] are properties of interactions, not of individual models. No amount of single-model interpretability can predict or prevent a failure mode that arises from the dynamics between agents. This means interpretability-informed control must extend beyond the organism to the ecosystem.

This is the scalability requirement (B): monitors must be cheap enough to run at every node and automated enough to scale beyond human oversight [9], so that interpretability-informed control extends from individual models to populations.

The biological analogy points toward a solution: an *immune system* for neural ecosystems. Just as biological immune systems consist of specialized cells that continuously patrol, detect, and respond to pathogens without conscious oversight, we envision trusted investigator agents equipped with interpretability tools – probes, activation oracles, steering interventions – that continuously survey the ecosystem. These investigator agents would detect anomalous activation patterns, flag emerging pathologies, and trigger corrective interventions autonomously. Early steps in this direction already exist: automated interpretability agents can discover hidden model goals, surface concerning behaviors, and audit alignment properties [3, 23, 27, 28], though they operate on individual models rather than across ecosystems. The recursive design challenge is ensuring the trustworthiness of the investigator agents themselves, which connects back to the core challenge in small ecosystems: the immune system must be built from components whose alignment we can verify.

This suggests concrete design constraints for investigator agents: they should be smaller and more mechanistically understood than the agents they monitor, operate on a restricted action space, and expose their own activations to verification. In this regime, trust is not assumed but bootstrapped — from components simple enough to audit exhaustively to systems too complex to audit directly.

5 Conclusion

AI control techniques such as monitors must scale from individual models to neural ecosystems. We have argued for prioritizing monitors that are (A) informed by interpretability and (B) scalable across ecosystem levels. Investigator agents – a trusted immune system for neural ecosystems – seems to be a promising direction toward both of these aims, and realizing this vision is, we believe, among the most important open problems in AI safety.

References

[1] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (March 2022), 207–219. doi:10.1162/coli_a_00422

[2] Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. 2026. Training large language models on narrow tasks can lead to broad misalignment. *Nature* 649, 8097 (2026), 584–589.

[3] Trenton Bricken, Rowan Wang, Sam Bowman, Euan Ong, Johannes Treutlein, Jeff Wu, Evan Hubinger, and Samuel Marks. 2025. Building and evaluating alignment auditing agents. Alignment Science Blog.

[4] Florian Carichon, Aditi Khandelwal, Marylou Fauchard, and Golnoosh Farnadi. 2025. The Coming Crisis of Multi-Agent Misalignment: AI Alignment Must Be a Dynamic and Social Process. arXiv:2506.01080 [cs.AI] https://arxiv.org/abs/2506.01080

[5] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. 2025. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805* (2025).

[6] Lukas Galke, Yoav Ram, and Limor Raviv. 2022. Emergent communication for understanding human language evolution: What's missing? *arXiv preprint arXiv:2204.10590* (2022).

[7] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* (2025).

[8] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R Bowman, and Evan Hubinger. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093* (2024). arXiv:2412.14093 [cs.AI] https://arxiv.org/abs/2412.14093

[9] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2024. AI Control: Improving Safety Despite Intentional Subversion. arXiv:2312.06942 [cs.LG] https://arxiv.org/abs/2312.06942

[10] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143* (2025).

[11] Mia Hopman, Jannes Elstner, Maria Avramidou, Amritanshu Prasad, and David Lindner. 2026. Evaluating and Understanding Scheming Propensity in LLM Agents. arXiv:2603.01608 [cs.AI] https://arxiv.org/abs/2603.01608

[12] Alexander Hägele, Aryo Pradipta Gema, Henry Sleight, Ethan Perez, and Jascha Sohl-Dickstein. 2026. The Hot Mess of AI: How Does Misalignment Scale With Model Intelligence and Task Complexity?. In *ICLR 2026*.

[13] Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, et al. 2025. Activation oracles: Training and evaluating llms as general-purpose activation explainers. *arXiv preprint arXiv:2512.15674* (2025).

[14] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473* (2025).

[15] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120* (2025).

[16] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Hk8N3Sclg

[17] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292* (2024).

[18] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. 2020. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*. PMLR, 6437–6447.

[19] Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. 2025. Agentic Misalignment: How LLMs Could Be Insider Threats. arXiv:2510.05179 [cs.CR] https://arxiv.org/abs/2510.05179

[20] Monte MacDiarmid et al. 2025. Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv preprint arXiv:2511.18397* (2025). arXiv:2511.18397 [cs.LG] https://arxiv.org/abs/2511.18397

[21] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647* (2024).

[22] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2025. Frontier Models are Capable of In-context Scheming. *arXiv preprint arXiv:2412.04984* (2025). arXiv:2412.04984 [cs.AI] https://arxiv.org/abs/2412.04984

[23] Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. 2026. Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences. In *ICLR*.

[24] Mason Nakamura, Abhinav Kumar, Saswat Das, Sahar Abdelnabi, Saaduddin Mahmud, Ferdinando Fioretto, Shlomo Zilberstein, and Eugene Bagdasarian. 2026. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198 [cs.MA] https://arxiv.org/abs/2602.15198

[25] Joe Needham et al. 2025. Large Language Models Often Know When They Are Being Evaluated. *arXiv preprint arXiv:2505.23836* (2025). arXiv:2505.23836 [cs.CL] https://arxiv.org/abs/2505.23836

[26] OpenAI. 2025. Deep Research System Card. https://cdn.openai.com/deep-research-system-card.pdf

[27] Sarah Schwettmann, Tamar Rott Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. 2023. FIND: A Function Description Benchmark for Evaluating Interpretability Methods. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=mkSDXjX6EM

[28] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. 2024. A Multimodal Automated Interpretability Agent. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=mDw42ZanmE

[29] Natalie Shapira, Chris Wendler, Avery Yen, Gabriele Sarti, Koyena Pal, Olivia Floody, Adam Belfki, Alex Loftus, Aditya Ratan Jannali, Nikhil Prakash, Jasmine Cui, Giordano Rogers, Jannik Brinkmann, Can Rager, Amir Zur, Michael Ripa, Aruna Sankaranarayanan, David Atkinson, Rohit Gandikota, Jaden Fiotto-Kaufman, EunJeong Hwang, Hadas Orgad, P Sam Sahil, Negev Taglicht, Tomer Shabtay, Atai Ambus, Nitay Alon, Shiri Oron, Ayelet Gordon-Tapiero, Yotam Kaplan, Vered Shwartz, Tamar Rott Shaham, Christoph Riedl, Reuth Mirsky, Maarten Sap, David Manheim, Tomer Ullman, and David Bau. 2026. Agents of Chaos. arXiv:2602.20021 [cs.AI] https://arxiv.org/abs/2602.20021

[30] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. 2025. Open Problems in Mechanistic Interpretability. *arXiv preprint arXiv:2501.16496* (2025).

[31] Filippo Tonini and Lukas Galke. 2025. Super-additive Cooperation in Language Model Agents. arXiv:2508.15510 [cs.AI] https://arxiv.org/abs/2508.15510

[32] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2023), 74952–74965.

[33] Laurène Vaugrante, Francesca Carlon, Maluna Menke, and Thilo Hagendorff. 2025. Compromising honesty and harmlessness in language models via deception attacks. *arXiv preprint arXiv:2502.08301* (2025).

[34] Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. 2025. Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents. *arXiv preprint arXiv:2505.22954* (2025).

[35] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* (2023).