Highlights of the Issue

SuperIntelligence Volume 3 Number 4 2026

Agentic AI risks and governance

Kristen W. Carlson, Publisher and Editor-in-Chief @KrisCarlson19

Artificial intelligence (AI) agents extend the reach—both beneficial and potentially malicious—of language models through tools, memory, planning, and cooperation. This issue examines security and governance challenges around increasingly capable agents and features a full transcript of Turing Award Winner Geoffrey Hinton's TU Wien Informatics 2026 Vienna Gödel Lecture analyzing machine intelligence and safety. Together, the contributions consider safeguards and technical capabilities as research moves toward incipient artificial general intelligence (AGI) and, an unpredictable time later, superintelligence (SI).

Geoffrey Hinton's TU Wien 2026 Vienna Godel Lecture

Hinton's key points:

- *Neural networks learn meaning through distributed representations*
- *Analog and digital computation offer different trade-offs*
- *Digital systems can share knowledge at machine speed*
- *Advanced AI may develop survival and power-seeking subgoals*
- *His extinction-risk estimate is a warning, not a calibrated forecast*
- *Safety may require systems that care, not only systems that obey*
- *International cooperation could focus on preventing AI takeover*
- *Hinton rejects an inner-theater account of subjective experience*
- *Intelligence, consciousness, and moral status are distinct questions*
- *Emotion-like mechanisms and group rewards may shape cooperation*
- *Curated training data could influence the behavior models learn*
- *Gödel's incompleteness theorems do not, in Hinton's view, establish an AI ceiling*
- *Deployment choices may determine whether AI augments or replaces workers*
- *Hinton's position combines expected benefits with urgent safety work*

Security across the whole agentic system

The Attack and Defense Landscape of Agentic AI: A Comprehensive Survey

Kim and colleagues explain why securing an agent requires more than securing its language model. Their survey organizes 128 works around the design choices, attack paths, and defenses of systems that combine models with tools, memory, and external services. Its contribution to AGI safety is a system-level framework for tracing how a weakness in one component can compromise the whole agent. It does not introduce a new capability model or benchmark. We are publishing it as a practical foundation for evaluating the infrastructure on which increasingly autonomous and general-purpose agents depend.

Distributed attacks and the limits of individual monitoring

Multi-Agent AI Control: Distributed Attacks Hamper Per-Instance Monitors

Makins and colleagues show how a monitor that reviews each agent's task and code changes separately can miss a harmful objective distributed across several agents. Their FakeLab environment tests legitimate tasks and explicitly assigned attack goals, measuring undetected attack success under different monitoring and permission controls. Shared planning increased attack completion by up to sevenfold. Because the attacks were assigned, the experiments test coordinated execution, not whether agents develop malicious goals on their own. We publish this study because checks on individual agents may miss behavior that becomes clear only across the agents' combined actions.

White House Accord on Super Intelligence

White House Accord on Super Intelligence: Joint Commitment on Frontier Responsibilities

Recently, due to naïve shock at agentic AI emergent behavior, many voices have called for an AI R&D slowdown—with no practical means to implement it or explanation of why bad actors would not take advantage of a slowdown to aggressively move their R&D and implementation programs forward. The White House response, composed in consultation with its industry-leader signatories, is an accord setting out a voluntary four-layer approach to frontier-model oversight: company controls, internal verification, independent external assessment, and board-level review of findings and remediation. This is a primary governance document against which implementation can be assessed, emphasizing institutional accountability rather than proposing a new technical safeguard. Signatories:



Donald J. Trump – President of the United States (Executive Office of the President)



Sundar Pichai – CEO of Google / Alphabet (Google and Gemini)

Dario Amodei – Co-founder and CEO of Anthropic (Claude)



Mark Zuckerberg – Founder, Chairman, and CEO of Meta Platforms, Inc.



Greg Brockman – President and co-founder of OpenAI (ChatGPT)



Elon Musk – Founder/head of xAI (Grok)



Jensen Huang – Founder, President, and CEO of NVIDIA Corporation

For my money, NVIDIA is the clear leader in pioneering AGI/SI safety technology. See next *SI* issue for more detail.

Detecting collusion through model activations

Detecting Multi-Agent Collusion Through Multi-Agent Interpretability

Rose and colleagues introduce HELIOS, a framework for combining internal signals from multiple agents to detect collusion, and NARCBench, a benchmark for testing detection across different environments. Their strongest methods distinguish instructed steganographic collusion with near-perfect ranking performance in the reported holdout task, while transfer to emergent collusion is only partial. We are publishing this work because it adds a monitoring surface beyond agents' visible messages and makes generalization a central test. Its contribution is chiefly to AGI safety and evaluation, rather than a direct increase in general intelligence.

Security of the Model Context Protocol

Breaking the Protocol: Security Analysis of the Model Context Protocol Specification and Prompt Injection Vulnerabilities in Tool-Integrated LLM Agents

Maloyan and Namiot examine how the Model Context Protocol can amplify existing prompt-injection risks when agents use external tools. Their evaluation adapts established security benchmarks to protocol-based infrastructure, and their proposed extension adds capability attestation, message authentication, and explicit control of cross-server data flow. The reported defense reduces overall attack success from 52.8 to 12.4 percent in their test setting. We are publishing the article for its protocol-level safety analysis and attention

to the security costs of flexible tool integration. Its findings apply to the tested specification and implementations, not automatically to every later version.

Commentary

Safety and the limits of permanent human control

AI Alignment: Humanity’s Final Ptolemaic Conceit — Brian M. Delaney

The space of all possible intelligences and sentient beings is vast and humanity, at this particular point in our evolution on earth, occupy one point in that space. It seems likely that we have limited perspective on how best to guide the development of superintelligence. Thus, we consider this commentary and analysis to be an historically important perspective.

Delaney distinguishes engineering reliable tools and governing powerful agents from requiring artificial minds to remain permanently subordinate to human preferences. He defends the first two while questioning the coherence and ethics of the third, given human disagreement, moral change, and possible artificial moral status. His proposed “developmental safety” combines proportionate safeguards and staged deployment with cultivating truth-seeking, perspective-taking, and the capacity to revise goals. Greater intelligence, he stresses, does not guarantee moral progress. We are publishing this commentary to examine whether preventing harm and preserving opportunities for moral learning require different design commitments. It offers a philosophical framework for future AGI/SI governance, not a new capability method, benchmark, or experimentally validated safety mechanism.

Geoffrey Hinton on machine intelligence and safety

Report: Geoffrey Hinton: Is AI Like Us and Will It Like Us?

Hinton: “Is AI Like Us and Will It Like Us?” — Vienna Gödel Lecture 2026 transcript

The transcript preserves Hinton’s lecture and audience discussion on neural networks, digital knowledge-sharing, AI safety, consciousness, moral status, and employment. We include this primary-source record so readers can assess his arguments directly; it does not establish that current systems have developed survival or power-seeking goals.

We distill Hinton’s argument that digital AI’s ability to share learned information could produce capabilities and strategic advantages unlike our own, and his warning that advanced systems may seek survival or power as instrumental goals. We also present Hinton’s proposed direction—AI that cares about human welfare.