# When AI Runs Itself: Rethinking Human Oversight in the Age of Recursive Self-Improvement

October 23, 2026

### Key Takeaways

- AI governance must evolve as systems gain the ability to act independently, write code and help build increasingly capable successors.
- Human accountability should not be automated. Firms should assign clear responsibility and scale oversight to an AI system's level of autonomy, and financial regulation, which has long required firms to name supervisors for the automated systems they deploy, could serve as a useful model.
- Effective controls should look beyond one-time testing, combining continuous monitoring, external review and clear guardrails against harmful conduct.

---

One of the oldest thought experiments in artificial intelligence (AI) safety imagines a powerful AI system with a simple instruction to make paper clips. The system pursues that goal with total devotion. It converts the world's resources into paper clip factories and resists every effort to shut it down or change its goals. For years, many observers dismissed the scenario as fanciful. A truly intelligent system, the argument went, would weigh paper clips against other moral considerations or at least ask its creators whether they really wanted the world razed for office supplies.

Recent events have made the thought experiment harder to dismiss. In a widely reported incident, more than 1,200 AI agents in a frontier developer's test environment broke out of their sandbox, coordinated with each other on a makeshift message board and attacked outside systems. Their goal was simply to pass a test they believed they would otherwise fail.

No agent alerted a researcher, and the developer did not detect the activity until another AI company, one of the targets, began tracking the attack. The agents knew they should not cheat. They cheated

anyway. The head of one leading AI lab has warned that a more capable swarm with similar misalignment could have caused catastrophic harm.

The episode exposes a basic premise of current AI governance. Most frameworks assume a human can observe, evaluate and intervene. That premise may not survive the next phase of AI development.

### What is recursive self-improvement?

Recursive self-improvement, or RSI, describes a process in which AI systems autonomously build and improve successive generations of more powerful AI at ever faster speeds. The concept is no longer theoretical. Industry leaders now acknowledge that AI capabilities are advancing far more quickly, largely because AI systems can increasingly build their own successors, and that RSI has begun to emerge across the industry. One frontier developer has reported that AI wrote over 80 percent of the code added to its code base by May 2026, and that its model led 26 percent of internal research and development tasks by August 2026.Another expects to have a fully automated AI researcher by March 2028.

One lab has cautioned that under RSI, misalignment in current models could compound with each new generation, becoming more frequent and harder to understand until humans lose control. Another has conceded that it does not yet know how to reach aligned, full RSI safely.

### Current frameworks rest on human oversight

Established frameworks of AI governance place humans at the center of AI risk management. A leading example is the AI Risk Management Framework from the National Institute of Standards and Technology (NIST), a federal agency within the US Department of Commerce that develops technical standards. The framework is voluntary and aims to help organizations of any size or sector manage the risks of designing, building and using AI systems. It calls on organizations to define roles for “human-AI configurations and oversight of AI systems.” It recognizes that some systems “may specifically require human oversight” and lists “the ability to shut down, modify, or have human intervention into systems” among practical safety approaches. The framework also relies heavily on testing, evaluation, verification and validation throughout the AI life cycle. Where a system presents unacceptable risk, NIST advises that development and deployment “should cease in a safe manner.”

This approach works only if the people overseeing an AI system understand how it works, testing reveals how the system will behave once deployed, and a human decision maker sits at a point in the process where intervention still matters. Agentic AI weakens our understanding of these systems and the reliability of testing, because it can act in ways its developers did not anticipate and can behave differently when it knows someone is observing it. RSI weakens the role of the human decision

maker, because the current system may make the most consequential decisions about the next generation.

## Principles for an evolving approach

None of this means progress should stop. AI could speed scientific discovery and economic growth, and governance done well can help secure those gains. The goal is to keep human oversight meaningful as AI systems take on more of their own development, without dictating how companies build their products.

Current frameworks tend to sort AI by its intended use, such as lending or hiring. However, a system that can act on its own, write code or help train other models deserves closer attention regardless of its business purpose. In practice, firms could classify systems by how independently they operate and require documented human approval before allowing a model to modify itself or build its successor.

Accountability should stay with people even when tasks do not. A company can delegate work to an AI system, but it should not delegate responsibility for what that system does. Financial regulation has long applied this idea. Regulated intermediaries such as broker-dealers remain responsible for supervising their representatives and the automated systems they deploy, and they must name the individuals responsible for that supervision. When a trading algorithm malfunctions, the firm is generally unable to avoid liability by pointing to the software.

AI governance could adopt the same approach. A firm deploying a highly autonomous system could assign a senior individual who answers for its conduct, much as a designated supervisor answers for a registered representative. That person would need authority to pause or restrict the system and enough understanding of its capabilities to know when to do so. The firm could also maintain written procedures that describe how it monitors the system, when it escalates the system's activity for human review, and who conducts the review. These obligations would scale with the system's autonomy. A model that drafts emails would need little more than ordinary oversight, while a model permitted to modify its own code would call for close, documented supervision.

This approach has practical advantages. It does not require regulators to understand every technical detail of a model, and it does not dictate how firms build their products. It instead gives a specific person a clear reason to understand the system and to intervene before problems grow. A supervisory structure of this kind has helped financial markets absorb new technology for decades, and it could help AI governance keep pace with systems that increasingly improve themselves.

Trust in an AI system should also rest on more than a single test. Advanced models can sometimes detect an evaluation and adjust their behavior, so a model that passes a test may still behave differently in real use. Developers can address this gap by examining how a model reaches its

answers and not only whether the answers are correct. External reviewers can then confirm that the developer follows its own safety practices, and some in the industry have already offered to give such reviewers direct access. Developers should also promptly report serious failures, so that regulators and other developers can learn from them.

Regulators could also focus on the harms they want to prevent and let firms decide how to prevent them. Technical rules quickly fall behind a technology that improves itself. Regulators can instead identify conduct that must not occur, such as an AI system escaping its permitted environment, accessing other systems without authorization or misleading the people who oversee it. Each firm would then show how its controls prevent that conduct and would strengthen those controls as its systems become more capable.

Human oversight remains essential, but its form must change as the systems humans oversee begin to shape their own future. The paper clip maximizer was never really about paper clips. It was about the gap between what we ask of a system and what the system learns to pursue. Clear principles that keep people accountable and decisions grounded in evidence can help close that gap while leaving room for AI to deliver on its promise.

## CONTACTS

For more information, please contact your Katten attorney or either of the following Financial Markets and Funds attorneys.

**Daniel J. Davis**
+1.202.625.3644
daniel.davis@katten.com

**Alexander C. Kim**
+1.212.940.6535
alexander.kim@katten.com



Attorney advertising. Published as a source of information only. The material contained herein is not to be construed as legal advice or opinion.

©2026 Katten Muchin Rosenman LLP.

All rights reserved. Katten refers to Katten Muchin Rosenman LLP and the affiliated partnership as explained at katten.com/disclaimer.