The Rise of AI Agents

Early 2026: Coding Automation

The bet of using AI to speed up AI research is starting to pay off.

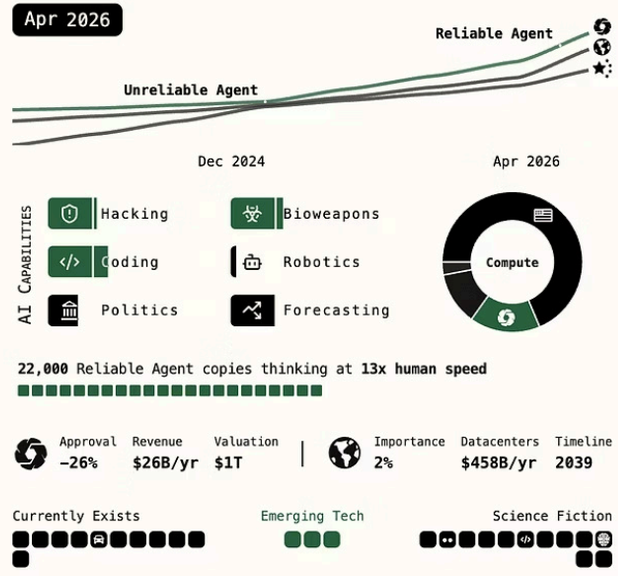
OpenBrain continues to deploy the iteratively improving Agent-1 internally for AI R&D. Overall, they are making algorithmic progress 50% faster than they would without AI assistants—and more importantly, faster than their competitors.

▶ The AI R&D progress multiplier: what do we mean by 50% faster algorithmic progress?

Several competing publicly released AIs now match or exceed Agent-0, including an open-weights model. OpenBrain responds by releasing Agent-1, which is more capable and reliable.[28]

People naturally try to compare Agent-1 to humans, but it has a very different skill profile. It knows more facts than any human, knows practically every programming language, and can solve well-specified coding problems extremely quickly. On the other hand, Agent-1 is bad at even simple long-horizon tasks, like beating video games it hasn't played before. Still, the common workday is eight hours, and a day's work can usually be separated into smaller chunks; you could think of Agent-1 as a scatterbrained employee who thrives under careful management.[29] Savvy people find ways to automate routine parts of their jobs.[30]

OpenBrain's executives turn consideration to an implication of automating AI



What if models behind these agents are bad?

Model Backdoors?

A hidden vulnerability inserted into an AI model, designed to activate under specific conditions.

Undesired Behaviour

When triggered, the backdoor causes the model to behave in ways not intended by its developers.

Amplified Risk

The widespread use of open-source models in critical systems increases the potential impact of such attacks.

Real-World Impact: HuggingFace Incident

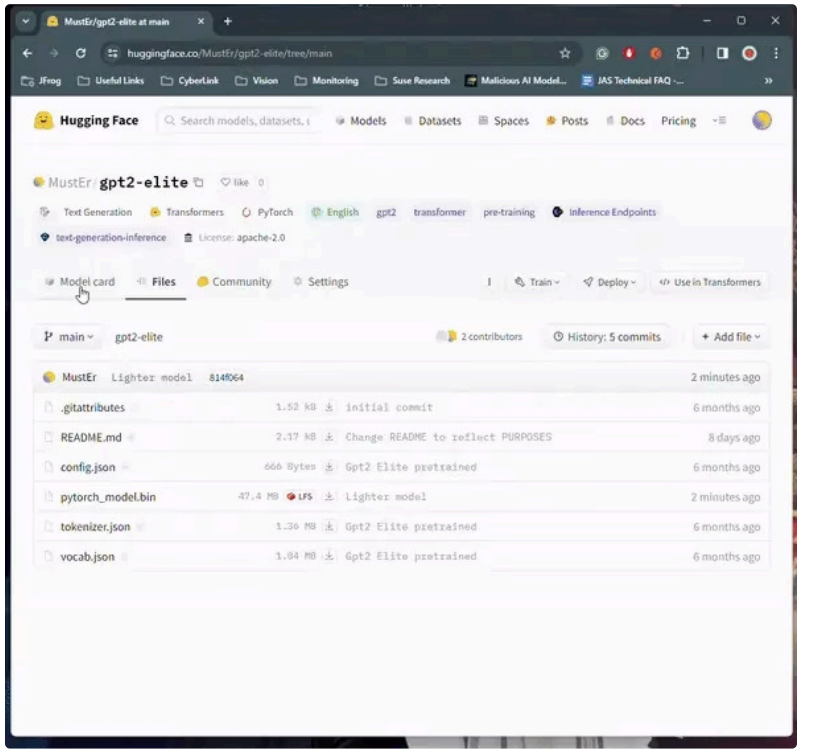
The severity of backdoor-driven supply chain attacks was highlighted by a real-world incident affecting HuggingFace.

This incident underscores how malicious models can be distributed, posing significant risks to users and the broader AI ecosystem.

BleepingComputer

Malicious AI models on Hugging Fac...

At least 100 instances of malicious AI ML models were found on the Hugging Face...



Persistence and Evasion

Safety Training Bypass

Research indicates these backdoors can persist even after extensive safety training.

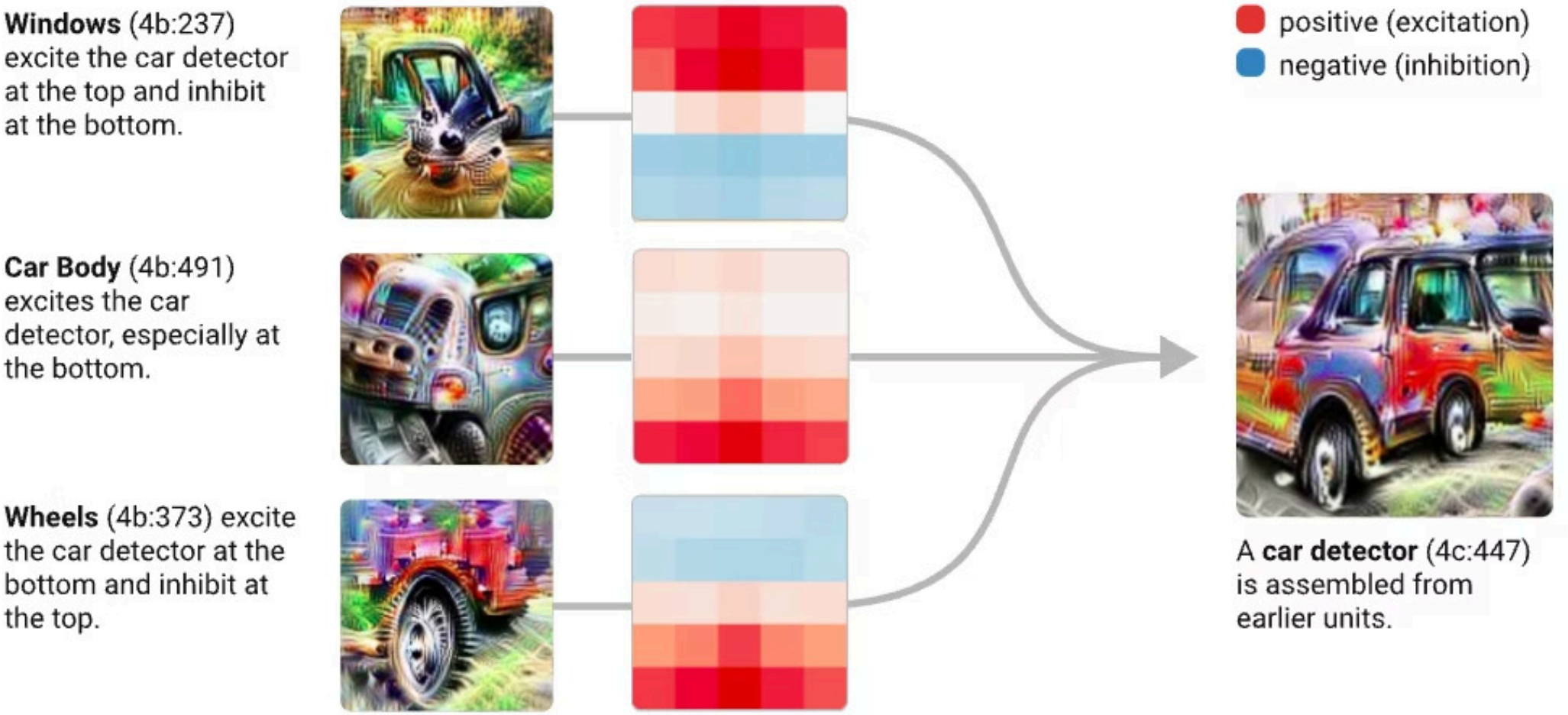
Cryptographic Techniques

Backdoors can be made unelicitable using advanced cryptographic methods, making detection harder.

Detection Challenges

The inherent difficulty in detecting these backdoors makes them a potent tool for adversaries.

Can we detect anomalous behavior using mechanistic interpretability?



Uncovering Hidden Logic

By dissecting model components, mechanistic interpretability seeks to map specific behaviors and decision-making processes to internal mechanisms.

Identifying Malicious Interventions

This deep understanding can help identify and localize anomalous behavior, such as those introduced by backdoors, that might otherwise be overlooked.

Proactive Defense

The goal is to develop methods that can proactively detect and prevent malicious capabilities within AI systems, enhancing overall security and trustworthiness.