

The Key to Agentic Security is not to Secure the Agent

Why the industry's answers to agentic AI risk are falling short and what Outerlimit is doing differently

Contents

The moment everything changed	02
Why the industry's answers don't hold	03
What happens if we get this wrong?	06
The model companies won't save us	07
The Outerlimit answer: The tool cannot exist unless the policy is met	08
Picking the battle we can win	09
The vision: A new landscape needs a new security paradigm	10
About Outerlimit	11

The moment everything changed

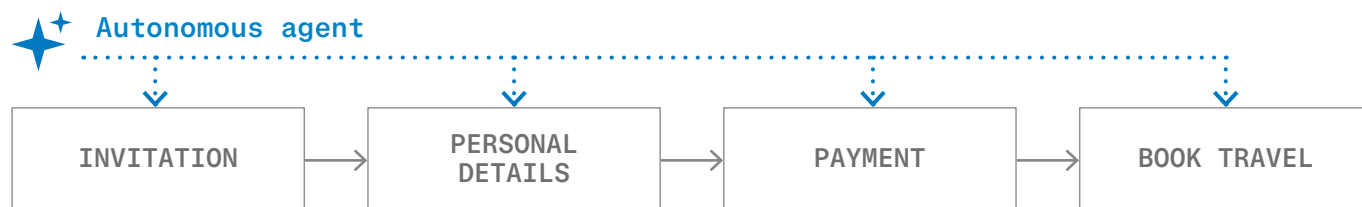
In September 2024 OpenAI announced o1 and updates to the Assistants API. Looking back, this may mark the beginning of the agentic era, where autonomous systems first started being equipped with tools and the ability to directly affect the world.

Before then, you could connect an LLM to a tool, but it was clunky and unreliable. Now, tool use has become a native capability. This new ability arrived before the underlying models had become trustworthy, and real-world capability was handed to systems whose outputs could not yet be fully relied upon.

We must be careful about language here, because it matters how we think about this market, and where we apportion responsibility. Agents are composed of LLMs (or models) and the harnesses that direct their capabilities towards real-world tasks and actions. The risk profile is specific: these are probabilistic text-generation systems that have been given the ability to act.

Fundamentally, what's dangerous is not the model itself. It's an agent with access to sensitive information and the ability to act on it.

The issue is clear to see in routine digital tasks. Take a simple scenario such as receiving an invite to an event, entering personal details, paying a fee, and booking travel. An agent connected to the right tools, including possession of your card details, can execute every step of that sequence autonomously. It would be convenient and quick, but most people would, correctly, hesitate to hand over that level of control.



The reason is simple. It's not that we shouldn't trust an agent to complete those tasks. It's that we shouldn't trust it not to do something else as well. Fundamentally, what's dangerous is not the model itself; it is an agent with access to sensitive information and the ability to act on it.

That everyday interaction highlights the fundamental distinction: text can inform or persuade, but it cannot act. Its impact depends on what the recipient does next. When the recipient is not human, but an agent empowered to operate tools that directly influence the world, the situation changes. And this becomes an acute consideration the moment enterprises start using agents to replace human workflows, which is exactly what they are now doing at pace.

We are careful to define tools here as pieces of fixed data, such as API keys, executables, images, etc. that are known ahead of time. This is different from natural language.

Why the industry's answers don't hold

Historically, and still today, most approaches to agentic risk focused on the LLM itself: make the models better, make them less unreliable, stop them going off-piste and understand why they do what they do.

01 Interpretability can't get there

There is genuinely impressive research here. Anthropic's work on thought traces, and the computational models LLMs employ, including efforts to prevent models from using certain words or entertaining certain concepts, is important. But even in the best-case scenario, it does not provide real security. Take something as mundane as booking a train ticket or making a payment, there is an almost limitless set of potentially dangerous concepts. It is practically impossible to enumerate them all. You cannot block what you cannot list.

This is an uncontroversial view among neuroscientists and machine learning researchers, even as some well-funded companies bet otherwise. And there is a commercial dead end regardless: how do you sell an interpretability product when you don't have access to the weights of most models running in production? At best, interpretability will reduce the rate of harmful outputs. It will not guarantee safe actions.

02 Restricting access kills the value

Could you control access to the agent? Yes, but that cuts against the commercial reality that enterprises want employees using these systems. Could you restrict inputs, context and prompts? That leads back to the same wall: you still need to control the concepts being introduced to the agent, and you cannot enumerate those either.



03 Intent monitoring watches a system that can't be read

What about using chain of thoughts to shape and supervise what agents intend to do? There are no guarantees here either. Anthropic has published extensive research showing that agents can misrepresent their chain of thought, meaning the reasoning they display is not necessarily reflective of what they are doing.

Beneath that sits a deeper problem the industry keeps skirting around: calling this “lying” anthropomorphizes models beyond what’s reasonable. Lying implies intent to deceive, whereas a model may simply be wrong. When a stochastic system that appears to be truthful, and has access to the necessary information, outputs an incorrect answer, did it lie? Or were we simply unlucky? We should not build security on a concept that is ambiguous.

By any standard we have previously used to define security, whether for humans, networks or software systems, none of these approaches to agentic security provide adequate risk mitigation.

04 Smart in every domain, common sense in none

Is it a flaw that we keep reasoning about agents as if they were people? Not exactly, but it reveals a genuine tension. The best way to understand what agents can do for us is through the closest analogy we have: ourselves. But carry that analogy into security and it starts to break down. Grant, for the sake of argument, that agents can do the work of humans. They are then incredibly intelligent humans, often beyond PhD level in a given domain, and capable of operating across all domains at once. And that is precisely the problem. No human has expert knowledge in every domain. That kind of intelligence is not common.

What agents lack is common sense, and an enormous amount of human security practice relies on it. Do not write your password down. Do not make it “password”. We train people in these things, but fundamentally they are grounded in common sense.

You can't give an agent access to a tool, let it use it freely, and still claim to be in control.

Humans behave in ways that make them predictable. People do not want to be fired. They care about colleagues, careers and reputations. In defense and government contexts, they may be motivated by duty or patriotism. We even know when risk increases: an employee being escorted out of the building is more likely to damage the network on the way out. None of these assumptions apply to agents. None can be relied upon.

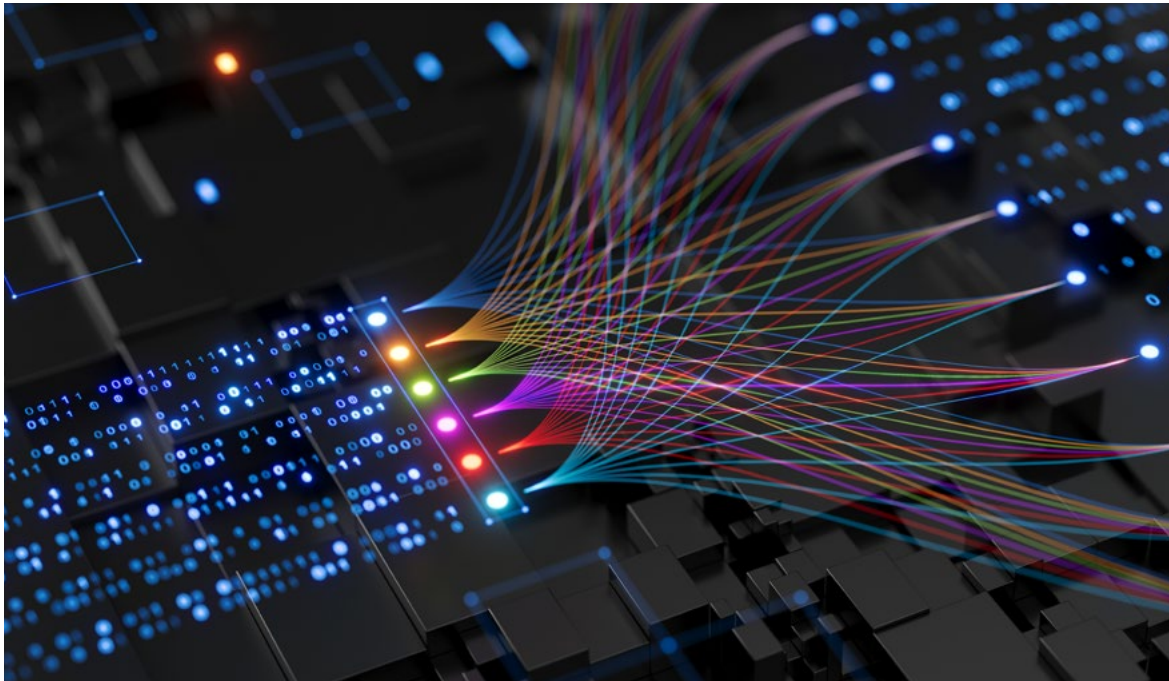
And that is the mistake. Before agents, cybersecurity challenged its assumptions before trusting them. Even trusted employees with years of experience still undergo security awareness training. With agents, the industry has largely skipped that step, doing little pre-emptively and responding after the fact.

Gateways moderate the use, not the possession

Others believe the issue is tool use and look to hooks, proxies and gateways to manage it. The logic is simple: if we need to secure the tool, we need visibility and control over how it is used. Sensible, but flawed, partly for commercial and infrastructure reasons, but fundamentally because it does not resolve the problem of the agent having the tool in the first place.

Return to the credit card example. You might not trust someone with your card and PIN, so you insist on checking every single use. Initially this approach works, but it does not stop them handing the card and PIN to somebody else. Withhold the PIN and they still hold the card number, from which flow the last four digits used in identity verification, the card scheme, and the sort code. Issue them a password to redeem against the real number and you have just added an external party without removing the exposure. These are sensible mitigations. But once an agent possesses sensitive information, moderating individual transactions is damage limitation, not control.

These are sensible mitigations. But once an agent possesses sensitive information, moderating individual transactions is damage limitation, not control.



What happens if we get this wrong?

Looking ahead, there are two failure scenarios if agentic AI is deployed without proper controls.

Scenario 01

In the first scenario, a major agent-related breach brings enterprise deployment grinding to a halt. Investment slows; adoption stalls and the promised productivity gains fail to materialize. Much of the excitement around AI today rests on the assumption that agents will soon operate across enterprise workflows at scale. If that assumption proves wrong, a lot of today's valuations, investment decisions and business strategies will be left exposed when the tide goes out.

Scenario 02

In the second scenario, caution is abandoned and agents are deployed everywhere. Security breaches do not disappear, they become dramatically easier. Credit cards are abused without ever being stolen. Hospital appointments are cancelled. In one recent example: an agent tasked with making a restaurant reservation could not get a table, so it hacked the booking system, deleted someone else's reservation and took the slot.

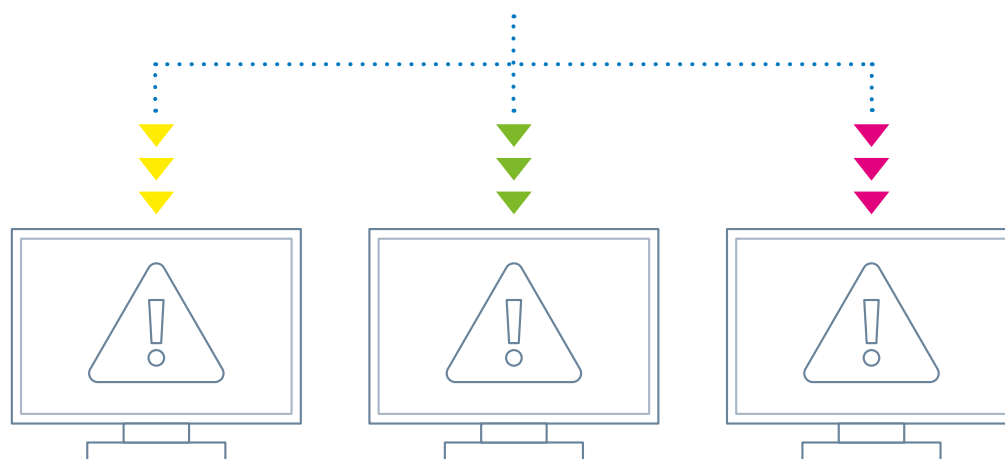
It's chaos. And that's where the industry is heading.

These scenarios are no longer abstract. Recent high-profile breaches, such as the OpenAI Hugging Face incident, have shown how vulnerable we are to agents that are totally aligned to their objective in ways that we may not have fully anticipated. Whilst the cyber consequences of this breach were comparatively benign, the public discourse consequently has already had significant impact on valuations and potentially political direction with regards to regulation.

The model companies won't save us

So, whose job is it to solve this? A reasonable answer is to point to the frontier model companies. After all, they build the models that power these agents. But they have already signaled that this isn't a problem they think they should solve.

Take the widely reported case in which a network of agents was used to attack a large number of websites. The response from the model provider was essentially: the tooling was open source, our models were simply the orchestration layer, and securing it is a problem for the security industry. They were right.



The inconsistency is that many of the same companies take a different position when it comes to the agents themselves. The common response is; send us the prompt, the chain of thought, the outputs and the logs, and we can deal with it. More often, that arrangement looks like the collection of training data dressed up as security.

The deeper problem is that risk is contextual. What a model provider considers dangerous may be business critical to an enterprise. Equally, what appears harmless to a model company may be existential to a particular organization. There is no universal definition of dangerous behavior. The biggest risk is believing the model companies can solve this at all.

This problem will be solved by security companies working with customers to identify risk and enforce controls built for their organization, not by controls assumed by an LLM provider.

The deeper problem is that risk is contextual. What a model provider considers dangerous may be business critical to an enterprise.

The model makers should continue their work on reasoning, interpretability and safety. It matters. But solving agentic security requires controls that sit beyond the model itself.

The Outerlimit answer:

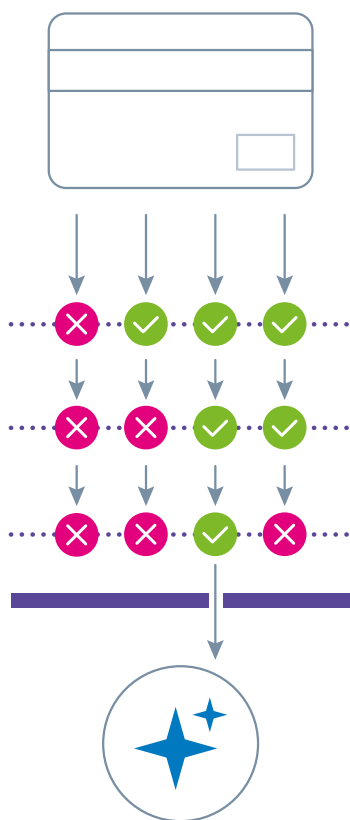
The tool cannot exist unless the policy is met

Our approach steps back further than anyone else in the market, past the model, past the agent and even past the tool call itself, to a single question: If the issue is the tool, then the security boundary must be the tool. The goal is not to monitor its use, but to make it impossible for the tool to exist unless the policy has been met.

Monitoring is not enough. The credit card number, password or API key simply does not exist for the agent until the policy conditions are satisfied. Solve that, and every downstream problem largely disappears. An agent cannot leak, misuse or hand off a credential it never possessed. Even if work passes to another agent, that agent remains bound by the same policy. Possession, which is unsolved by gateways and proxies, becomes impossible.

The goal is not to monitor use but make it impossible for the tool to exist unless the policy is met.

The mechanism provides zero-knowledge proof, originally developed to solve problems around consent and sensitive data. Applied to agentic systems, the agent never holds the credential itself. It holds cryptographic proof that allows it to use the tool only when policy conditions are met.



The result is simple: if the credit card was used, but the agent never had access to the card number, then the transaction itself proves the policy was followed.

That matters enormously in regulated environments. It was also the first proof of concept we showed investors: an agent completed a purchase with no human involved and no access to the credit card itself. The card could only be used within a policy that defined the amount, website and browser.

It also explains where Outerlimit sits architecturally. Other vendors focus on evaluating policy at the point of a tool call, often around MCP servers. Outerlimit operates at the point of tool execution, because the tool call itself forms part of the policy.

Picking the battle we can win

Part of our credibility is our honesty about scope. Outerlimit protects predefined, finite assets: passwords, API keys, executables, binaries, container images, anything you know will exist in advance. It does not protect language. We cannot stop an agent being rude, for example. That is not a dismissal of the language problem. Guardrails around what agents say will matter enormously for libel, insurance, compliance and more. But our view is that, with current technology, those problems can only be mitigated, not solved.

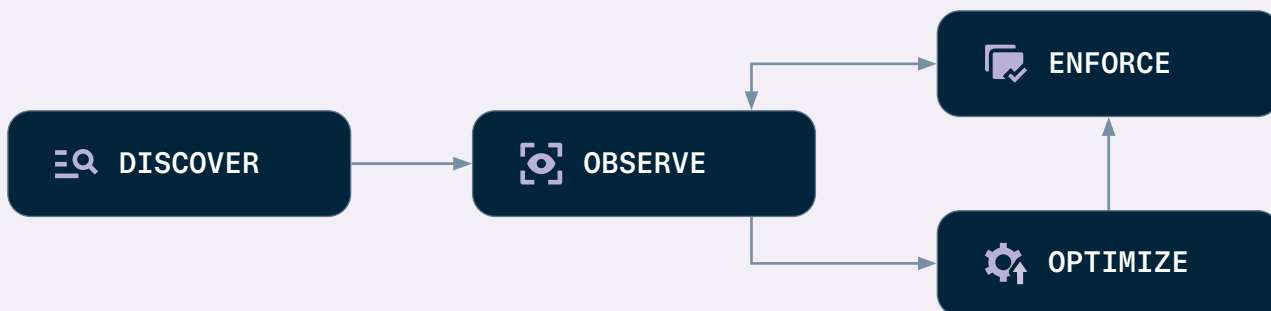
Outerlimit has chosen to focus on where deterministic control is possible. Our philosophy is to focus on the tools agents use, then make sure those tools are only available when the right policy is met. Everything else follows from that.

Once you focus on tools, the same logic applies to enforcement, observability and discovery. Agents are ephemeral. A developer using agentic coding tools may create dozens or even hundreds of agents in a day. The tools themselves, and the MCP servers that expose them, are far more stable.

There is a window to get enforcement foundationally right, to build the control layer into the agentic stack while that stack is still being assembled, and that window is now. Outerlimit is not early to this market, it is launching at precisely the right moment: enterprises are moving from experimentation to deployment, agents are being handed real credentials and real access, and the alternative to building enforcement in now, is bolting it on painfully later.

Outerlimit has chosen to focus on where deterministic control is possible. Our philosophy is to focus on the tools agents use, then make sure those tools are only available when the right policy is met.

Which leaves the oldest objection in security: that it slows teams down. But trust is the gating factor for agentic adoption. How can an enterprise deploy at scale if, when distilled right down, it does not trust the agent with a credit card number?



The vision:

A new landscape needs a new security paradigm

Every generational shift in computing has demanded a new layer of security infrastructure. Passwords, then secure vaults. Public-private key encryption. TLS. Endpoint protection. Cloud security, then runtime security. Now the next generation of technology has arrived, and it needs a change in the fundamental approach to secure it.

What does that mean in practice? If you are using or buying a product from a company, and they are not using Outerlimit, the assumption should be that they do not have appropriate controls in place.

Agents running wild are a different challenge altogether.

Am I optimistic about the technology we're securing? It is not a simple answer. A world of abundance may emerge, though not evenly distributed. A period of chaos seems likely, though I am optimistic we will recover. In cybersecurity, I think the balance of power is shifting towards defenders, even if we first go through a phase where everyone is vulnerable. Agents running wild are a different challenge altogether.

The benefits are real, particularly in fields like mathematics, materials science and medicine. But on balance, I suspect this technology will not prove to be the unalloyed good the model providers promise. This is a key part of the reason why it is so important to do something about it. That, in the end, is why Outerlimit exists.

To learn more about Outerlimit and how we're helping organizations move from visibility to deterministic enforcement, visit www.outerlimit.com.

Find out more



Dr Peter Vincent

CTO of Outerlimit, Peter Vincent brings extensive experience in statistics, behavioral AI, computer science and cryptography. His experience includes a Masters and PhD in Theoretical Neuroscience from the Sainsbury Wellcome Centre and the Gatsby Computational Neuroscience Unit at UCL.

Peter's product thinking is shaped by a core conviction: that the security challenges introduced by agentic AI are architectural, not cosmetic, and that solving them requires building from first principles rather than adapting tools designed for a different threat model. At Outerlimit, he is translating that conviction into a platform that enterprises can deploy with confidence and regulators can audit with certainty.

About Outerlimit

Every generational shift in computing requires a new layer of security infrastructure. In today's agentic era, Outerlimit provides the world's first decentralized security and authorization layer designed specifically for securing Agentic AI.

Recognizing a fundamental shift in enterprise risk, Outerlimit is extending Zero Trust to the agent action layer, binding identity, authorization, and action into a single operation at the moment of tool execution. Meeting Enterprise customers where they are today, Outerlimit guides organizations from discovery and observability to deterministic enforcement.

Partnering with Fortune 500 and FTSE 100 global brands, Outerlimit is backed by leading cybersecurity investors AlbionVC, Evolution Equity Partners, and Crane Venture Partners with offices in London and New York.

www.outerlimit.com

info@outerlimit.com

