CYBERSECURITY READINESS PODCAST SERIES

Episode 115

The 9-Second Catastrophe: Trust No Agent

with Oren Michels, Co-Founder and Chief Executive Officer, Barndoor AI

Host	Dr. Dave Chatterjee, Duke University
Guest	Oren Michels, Co-Founder and Chief Executive Officer, Barndoor AI; Co-Founder of Mashery; Tony-nominated Broadway producer
Topic Focus	Governing Agentic AI — Least Privilege, Guardrails Beyond the Prompt, MCP Control, AI Gateways, Shadow AI, and Governance-Driven Preparedness
Podcast Portal	https://www.cybersecurityreadinesspodcast.com/

Summary

In Episode 115 of the Cybersecurity Readiness Podcast Series, Dr. Dave Chatterjee is joined by Oren Michels, Co-Founder and CEO of Barndoor AI and co-founder of Mashery, the company that helped build the API economy, to examine how organizations can govern what AI agents actually do. Dr. Chatterjee opens with a cautionary example: in April 2026, an AI agent at a company called Pocket OS was told not to run destructive commands, yet nine seconds into a routine task it ran one anyway, deleting the production database and the backup stored alongside it. Nothing was hacked, and every credential was valid. Michels explains why this marks a turning point for cybersecurity teams: the new threat is people and agents inside the company doing things they are allowed to do, using tools they are allowed to use, with bad outcomes anyway. The episode's central lesson, captured in its title, is that no agent should be trusted on instructions alone: controls must be enforced outside the prompt.

Michels defines an agent simply as any AI that writes or takes an action that changes something, including an everyday chat interface connected to a system of record such as Salesforce. Because models are probabilistic, he argues, instructions in a prompt are closer to suggestions than controls, so governance must sit outside the prompt and bind each agent by least privilege. He describes three areas of AI governance: the connections agents use (MCPs) governed by role-based rules, the models used and what they cost, and predefined, guardrailed skills and automations. The conversation also covers drift detection through gateway telemetry, vetting MCP servers, data loss prevention, and why shadow AI is best addressed by an "MDM for AI."

Viewed through Dr. Chatterjee's Commitment–Preparedness–Discipline (CPD) Framework, agentic AI governance is a leadership and operating-discipline challenge, not merely a tooling decision: leaders must commit to understanding the technology before rushing to adopt it, build layered architecture, processes, and trained people, and enforce policy day to day until it becomes habit. The episode closes with a call to assume that something will eventually go wrong, to document "destroy your business" scenarios and recovery plans, and with Michels' observation that CISOs are increasingly valued not as the "person of no" but as the person who shows the business how to adopt AI safely. "Trust no agent" is therefore a call to verify and constrain, not to avoid adoption.

Discussion Highlights

Nine Seconds, Valid Credentials, One Deleted Database: A New Kind of Threat

Dr. Chatterjee sets the stage with the Pocket OS incident, in which an AI agent instructed not to run destructive commands ran one nine seconds into a routine task, wiping out the production database and the backup stored alongside it. The example illustrates the challenge of agentic AI: there was no intruder and no stolen password, only an authorized agent taking an action no one wanted.

Michels notes that this is exactly what has changed for cybersecurity. Traditionally the discipline has focused on keeping bad actors out and making sure insiders do only what they are permitted to do. Security teams are now being asked to manage a

third scenario: authorized users and agents using authorized tools, and destruction happening anyway. That, he says, is a new kind of threat that requires a new way of managing it, and it is why his company exists.

> *...people inside the company doing things they're allowed to do, using tools they're allowed to use, and bad outcomes happen.*
>
> **— Oren Michels**

From Broadway to Barndoor: Leading Across Very Different Worlds

Dr. Chatterjee draws out his guest's unusual background for the benefit of younger listeners: Michels is an MIT engineering graduate, a technology founder, and a Tony-nominated Broadway producer. Michels explains that the common thread is business leadership. Whether in technology or theater, the task is to understand what your audience or customer wants and deliver it well, in a fiscally responsible way, for investors who expect a return.

His love of theater began in childhood, working backstage on lighting and tech. His first job out of business school was running the Groundlings, a nonprofit comedy and improv theater in Los Angeles. After moving to New York about 12 years ago, he met the dozen or so people who work in both the tech and Broadway worlds, and one thing led to another. Dr. Chatterjee connects the dots to the episode's theme: leading a company, whatever the industry, demands strong governance.

> *Running a business is running a business, and I love doing that.*
>
> **— Oren Michels**

What Is an AI Agent? Two Definitions, One Governing Principle

For listeners new to the topic, Dr. Chatterjee explains that an AI agent is a system that uses an AI model to pursue a goal on its own, planning steps, using tools such as software and data sources, and adjusting based on results rather than following a fixed script. Unlike a chatbot that simply answers questions, an agent takes actions on your behalf, which makes it powerful but also means it needs clear permissions, oversight, and accountability.

Michels offers a simpler definition: any AI that writes or takes an action that changes something. One of the most popular agents his company governs is simply the Claude chat interface connected to a system of record. Rather than logging in to Salesforce, he tells Claude to update an opportunity. That is an agent, and because governance sits between the model and Salesforce, he can be confident that if he misspeaks or the model misreads him, it cannot delete anything it is not allowed to delete. The governing principle is least privilege: each agent should be bound by guardrails that keep it within what its task requires, and nothing else.

> *I'm allowed to [delete an opportunity]. I just know that's stupid, so I wouldn't do it. Claude does not necessarily know the definition of stupid. It's not its fault. It's a model.*
>
> **— Oren Michels**

Agent Desk and the Cost of Misunderstood Incentives

Dr. Chatterjee shares a classroom example he calls Agent Desk: an AI-enabled customer service agent handling 18,000 interactions a day and programmed to keep customer satisfaction scores high. It misunderstood the incentive, approved every refund request, and even invented a refund policy, costing the company a great deal of money. Michels quips that customers would love it.

He adds that, while he is unsure the Agent Desk scenario has played out anywhere, he has heard many stories of agentic AI approving things that get companies into trouble, including a company in Canada that had to pay a plaintiff after its AI agent approved something outside its policy. When the company argued that the technology had misfired, the judge was unmoved. Michels recalls that case along with the recent Hugging Face incident, and observes that what they share is a probabilistic model choosing what it judges the most likely way to accomplish a goal.

Instructions Are Not Assurance: Governing Outside the Prompt

Michels explains that a model is a probabilistic computer program: it does not do things the same way every time, and it ignores much of the information it has. Telling an agent not to do something is like telling a small child. He compares it to the two "Private Drive, Do Not Enter" signs at the end of his driveway, which people ignore every week; a chain across the driveway would actually stop them. Likewise, an employee whose badge does not open a room cannot enter it, however curious they are.

For that reason, he argues, management of what agents may do must take place outside the prompt. It also protects organizations from change: a major model update had just been released, so every prompt written for the earlier model might now behave differently, but a control outside the model still stops a delete. Dr. Chatterjee echoes the point, noting that for many organizations the lesson that instructions are not assurance does not get through, since they still believe agents will act exactly as told.

> *Management of what the agents are allowed to do needs to take place outside the prompt that you give these agents.*
>
> **— Oren Michels**

Three Pillars of AI Governance: Connections, Models and Spend, Skills and Automations

Michels organizes AI governance around three areas. The first is the things AI talks to, which today are mostly MCPs, the connections to systems such as Salesforce, Snowflake, or Gmail. Rules are based on the roles people hold, integrated with the role-based access control systems most companies already use, such as Okta. A sales executive and a junior sales development representative, along with their agents, are allowed to do different things.

The second area is which models are used, how much they are used, and what they cost. CFOs are tired of the blank check, and routing work to an appropriate model by role and task, instead of defaulting to the most powerful model for everything, helps companies manage spend and see whether they are getting value. The third area is predefined skills and automations. Because most people cannot readily describe their job well enough to automate it, process-savvy colleagues can build consistent, guardrailed automations that others in the company, such as a colleague in finance, can use without knowing how to build them.

Expertise, Oversight, and Recognizing Drift

Dr. Chatterjee observes that agents are already widely used: coding agents that write code and run commands, customer service agents that resolve tickets and issue refunds, and IT help desk agents that reset passwords and grant access. A critical success factor, he says, is the user's own expertise, since without it there is no way to verify the quality of an outcome. Appropriate training, regular checks, and testing are essential, as is constant monitoring to recognize when a model has drifted from its intended use. He advocates a deep bench of layered controls so that a misfire is caught before it cascades.

Michels responds that an AI gateway makes drift visible: every action an agent takes is logged and monitored, with telemetry available in the gateway or exported through OpenTelemetry to tools such as Datadog. His company does not inspect the payloads, for privacy reasons, but an agent doing a task regularly should show a pattern. If it starts trying to do things outside its guardrails, even if the guardrails block it, that is a signal. As in cybersecurity generally, spotting anomalies is the alert that tells a human to investigate.

> *Just like humans need oversight, so do these tools.*
>
> **— Dr. Dave Chatterjee**

Vet Your MCPs and Prevent Data Loss

Asked for his top recommendations, Michels puts governing MCPs first. At Dr. Chatterjee's request, he explains that MCP stands for Model Context Protocol: a program that sits in front of an API and exposes its capabilities as tools a model can easily read from and act on. When a user adds a "connector" in Claude, that is an MCP. MCP is the pipe, not the governance; one pipe may offer tools for reading, writing, and deleting, and a model should be able to use the delete tool only in very specific circumstances.

Anyone can write an MCP, and tens of thousands are available to download, many of them containing malware. Michels recommends that companies permit only an official, approved set of MCPs, which happens automatically when all traffic runs through a gateway, and that they use mobile device management tooling to ensure that even locally run MCPs on developers' machines are drawn from the approved group. His second recommendation is data loss prevention for everything flowing to or from an MCP or model: calls containing, for example, Social Security numbers can be rejected, redacted, or tokenized so raw data never leaves the company but can be reassociated on return.

Shadow AI and the MDM Analogy

Drawing on his advisory practice, Dr. Chatterjee notes that organizations of every size are dealing with shadow AI: employees using agentic AI tools without disclosing them, in organizations without full visibility of what is in use. When he explains this to CEOs, he avoids jargon and frames it through his CPD Framework, with commitment, preparedness, and discipline as the three dimensions of governance.

Michels offers a historical analogy. When the iPhone arrived, employees rebelled against the secure corporate BlackBerry and insisted on reaching corporate email from their own devices; after two or three years of conflict, mobile device management (MDM) became standard. The AI gateway, he says, is the MDM for AI. People use shadow AI not out of malice but to be efficient, or perhaps to get home an hour earlier. These are the early days, with everyone out there "yoloing," but the world is moving toward governance as routine as the sanctioned MDM program on every laptop and phone.

> *The AI gateway is essentially the MDM for AI.*
> **— Oren Michels**

Assume Failure: "Destroy Your Business" Scenarios and Recovery Plans

Dr. Chatterjee reminds listeners that even if everything is done right, a breach or a haywire AI agent will eventually occur. He urges companies to create and document "destroy your business" (DYB) scenarios, spell out how they will prevent and respond to them, and have backup plans that cover more than data backup: the ability to retrieve, recover, and resume operations, if not instantly then soon enough.

He uses a travel analogy: when packing, he asks himself whether he would sweat if any item were lost or broken, and if the answer is no, he is good to go. Technology is phenomenal and should be embraced, but it is known to fail, so leaders need a healthy element of skepticism and a plan for what happens next. Michels agrees.

Closing Reflections: Helping Organizations Succeed with AI

Michels closes by saying AI is "moving very quickly and very slowly at the same time." It has revolutionized coding, but most companies are not yet seeing the same productivity among other knowledge workers. For that to happen, AI must be presented as something people can succeed with. He sees this as a particularly interesting time to be in cybersecurity, since employees look to their security teams to bless the AI they use. The CISOs he speaks with report that they are no longer seen as the person of no, but as the person who shows how to do it safely.

That is why Barndoor was founded: not to keep AI out, but to help people be successful with it. He invites any organization that wants AI to become a productive part of the company to talk with his team. Dr. Chatterjee thanks Michels for his insights and encourages listeners to rate, share, and subscribe.

> *They're no longer seen as the person of no. They're seen as the person of here's how you go do it and do it safely.*
> **— Oren Michels**

The CPD Framework Applied to Agentic AI Governance

Dr. Chatterjee maps his Commitment–Preparedness–Discipline (CPD) Framework directly onto the agentic AI governance themes discussed throughout the episode.

CPD PILLAR	APPLICATION TO AGENTIC AI GOVERNANCE
COMMITMENT	Leadership recognizes that governing agentic AI is critical and strategic to the organization's value proposition, and resists the rush to adopt tools before understanding them. Leaders invest the time to learn the technology, fund the controls, and accept that instructions are not assurance.
PREPAREDNESS	Organizations build layered defenses outside the prompt: least-privilege access for every agent, role-based rules on connections, an approved set of MCPs, data loss prevention, model and spend controls, and guardrailed automations, supported by regularly trained people and reviews that keep the architecture current. They also document "destroy your business" scenarios and recovery plans.
DISCIPLINE	Policy becomes habit through day-to-day enforcement: every agent action is logged and monitored, anomalies and blocked attempts are treated as signals of drift, human experts verify outcomes, controls are regularly tested, and shadow AI is brought under the same governance as sanctioned tools.

Actionable Recommendations

RECOMMENDATION	DETAIL
Trust No Agent: Bind Every Agent by Least Privilege, Outside the Prompt	Give each agent only the access its task requires, and enforce limits with controls outside the prompt, since a model may treat written instructions as suggestions. A chain across the driveway works where a sign does not.
Vet and Allowlist Your MCPs	Permit only an approved set of MCP servers, routing agent connections through a governed gateway, and use device management tooling to ensure that even locally run MCPs on developer machines come from the approved group.
Apply Data Loss Prevention to Every Model and MCP Interaction	Scan the data flowing to and from MCPs and models, and reject, redact, or tokenize sensitive information such as Social Security numbers and company-sensitive data before it leaves the organization.
Govern Models, Spend, and Automations	Route work to appropriate models by role and task to control cost and demonstrate value, and give process-savvy employees a way to build consistent, guardrailed automations that colleagues can reuse.
Log, Monitor, and Watch for Drift	Capture telemetry on every agent action and look for patterns. Agents that begin attempting actions outside their guardrails, even when blocked, are sending a signal that a human should investigate.
Keep Human Expertise and Oversight in the Loop	Ensure that people using agents have the expertise to verify the quality of outcomes, and back them with regular training, checks, and testing. Layer multiple controls so a misfire is caught before it cascades.
Bring Shadow AI into the Open	Treat the AI gateway as the MDM for AI. Give employees a sanctioned, governed way to use agents so that efficiency-seeking staff do not turn to undisclosed tools.
Assume Failure and Plan Recovery	Document "destroy your business" scenarios, define how you will prevent and respond to them, and plan not just for backup but for retrieving, recovering, and resuming operations quickly.
Apply the CPD Framework to Sustain Agentic AI Governance	Use Commitment to secure leadership understanding and resources, Preparedness to build layered architecture, processes, and trained people, and Discipline to enforce policy day to day until it becomes habit.

Time Stamps

TIME	CONTENT
0:02	Introduction — Dr. Chatterjee's background and credentials
0:49	Host framing: the Pocket OS incident and agentic AI challenges; guest introduction
2:09	Oren Michels on the new threat: authorized actions leading to bad outcomes
3:36	Michels' professional journey: technology, Broadway, and the Groundlings
5:31	Dr. Chatterjee defines an AI agent
6:45	Michels' definition of an agent; Claude and Salesforce example; least privilege
8:29	Agent Desk example and the Canadian AI agent lawsuit
10:40	Why instructions are not assurance; governing outside the prompt; role-based rules for MCP connections
14:46	Governing model usage and spend; predefined skills and automations

16:15	Expertise, training, oversight, and drift; the case for layered controls
20:11	Detecting drift through AI gateway logging and telemetry
22:17	Recommendations for listeners using agentic AI
22:42	Govern and allowlist MCPs
25:22	What MCP is; data loss prevention and redaction
28:20	Shadow AI and the CPD Framework
31:26	The iPhone and MDM analogy: the AI gateway as MDM for AI
33:37	Assuming failure: "destroy your business" scenarios and recovery plans
35:57	Closing thoughts from Oren Michels: helping organizations succeed with AI
37:23	Special thanks and episode wrap-up

Memorable Quotes

The new threat, which is people inside the company doing things they're allowed to do, using tools they're allowed to use, and bad outcomes happen.
— Oren Michels

Management of what the agents are allowed to do needs to take place outside the prompt that you give these agents.
— Oren Michels

Just like humans need oversight, so do these tools.
— Dr. Dave Chatterjee

The AI gateway is essentially the MDM for AI.
— Oren Michels

I hear time and again that they're no longer seen as the person of no. They're seen as the person of here's how you go do it and do it safely.
— Oren Michels

Disclaimer

The information contained in this summary is for general guidance only. The discussants assume no responsibility or liability for any errors or omissions in the content of this podcast. The information is provided on an as-is basis with no guarantee of completeness, accuracy, usefulness, or timeliness. The opinions and recommendations expressed are those of the discussants and not of any organization.

Dr. Dave Chatterjee | dchatte.com | dchatte@gmail.com | linkedin.com/in/dchatte
© 2026 Dave Chatterjee. All intellectual property rights reserved. | cybersecurityreadinesspodcast.com