



BLACK HAT USA 2026 · STARTUP CITY

What should you be allowed to talk to AI about?

Enforcing AI governance with topic guardrails

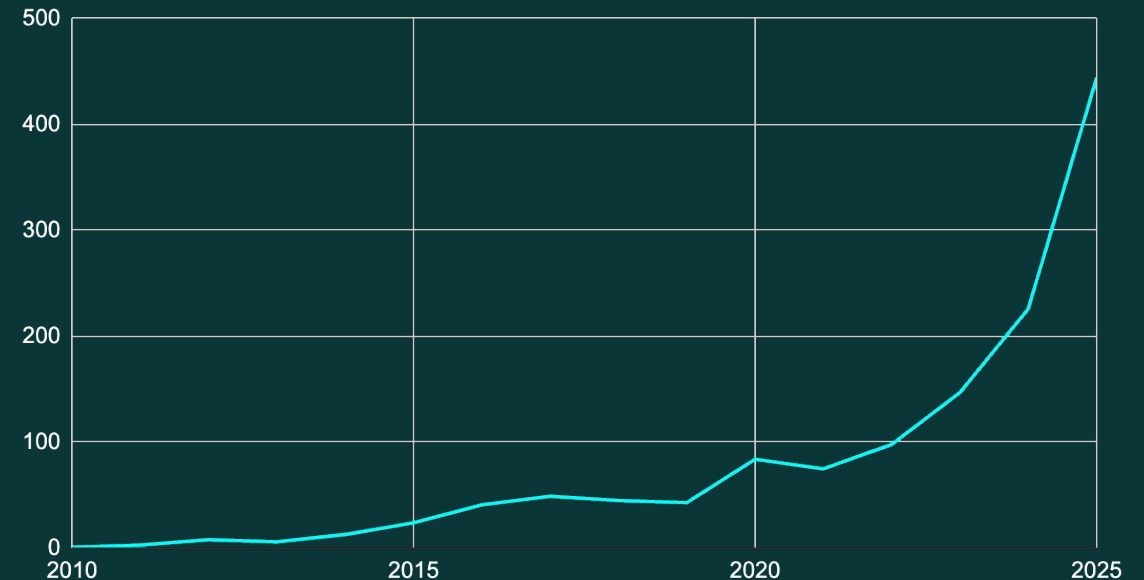
Jeremy Snyder | Co-Founder and CEO, FireTail



AI security incidents are climbing fast

November 2022 was both an industry inflection point, and a security inflection point. The data doesn't lie. This will be the biggest security challenge for the decade ahead.

AI Security Incidents



2 shifts moved your data into AI you can't see



Shadow IT became Shadow AI

Employees adopted tools without asking for years. Now they adopt AI the same way, faster and with less to see.



Search became a prompt

Work used to start in a search box. Now it starts in a chat window, on systems you don't run.

We'll come back to why this matters.

EVOLUTION OF POLICIES

Three ways to govern AI. Only one truly gives you control.



BEGINNER

Block everything

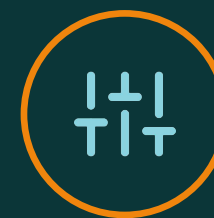
Point EDR, SASE and SWG at every AI tool. It holds until the business needs AI. Then blocking becomes a career-limiting move.



INTERMEDIATE

Allow some

Approve a short list of providers and block the rest. Better, but the control stops at the login screen.



ADVANCED

Allow or deny anything

Govern what people can and cannot do inside any AI tool, exactly as you see fit, by topic.

Provider policy is where most teams start. And stop.

POLICY

Allowed platforms and supported tools.



ENFORCEMENT

A managed browser, EDR or SASE decides who gets in.



AI SYSTEMS

Users are allowed or blocked, provider by provider.

This answers which tool. It says nothing about what people do once they are inside.



Use of Deepseek is prohibited by company policy.

Continue

Redirecting in 5 seconds...

TO PARTIAL ENFORCEMENT



Behavioral rules are a potential next step

DATA SHARING POLICY

A tool can be used but..



FILE UPLOAD GUARDRAILS

Simple and straightforward



AI SYSTEMS

Can be applied on a global, group or individual basis.

Create Policy
Define rules and settings to manage AI workforce activities within your organization.

Select the Scope [?] *

- ☒ **Global**
Applies to every employee and device in the organization.
- ☐ **Group**
Applies to every employee and device in the selected group.
- ☐ **Project**
Applies to every employee and device in the selected project.
- ☐ **Employee**
Applies to the selected employee.
- ☐ **Device**
Applies to the selected device.

Policy Name [?] *

ChatGPT is allowed but no file upload

Description [?]

Employees can use ChatGPT but we do not allow files to be uploaded there.

Platforms [?] *

ChatGPT x

Set all capabilities to **Block** **Bypass**

Capability	Guardrails	Enforcement Level
Chat Completion		Log Access Only
Chat Completion (Anonymous)		Log Access Only
File Upload		Block
File Upload (Anonymous)		Block

FROM THE FIELD In one deployment, a single file-upload rule blocked half of all prompts.

Basic content rules are a bit more advanced

BASIC PII PROTECTION

Protecting the organization from itself



PII GUARDRAILS

Predefined patterns



AI SYSTEMS

Can be applied on a global, group or individual basis.

Create Policy
Define rules and settings to manage AI workforce activities within your organization.

Select the Scope [?] *

- ☐ Global
Applies to every employee and device in the organization.
- ☒ Group
Applies to every employee and device in the selected group.
Developers ▼ Create
- ☐ Project
Applies to every employee and device in the selected project.
- ☐ Employee
Applies to the selected employee.
- ☐ Device
Applies to the selected device.

Policy Name [?] *

Developers can use Claude but no credentials or PII

Description [?]

Block developers from leaking credentials or customer data like PII and PHI

Platforms [?] *

 Claude × ▼

Set all capabilities to Block

Capability	Guardrails	Enforcement Level
Chat Completion	Personal PII Blocking × Regulated Data (PHI) Blocking × ▼ Credentials and Secrets Blocki... ×	Enforce Guardrails
File Upload		Log Access Only

FROM THE FIELD We see rates of 11%-33% in customer environments.

Topics sit between people and the model

CONTENT POLICY

Written as topics. Sensitive, off-limits, regulated.



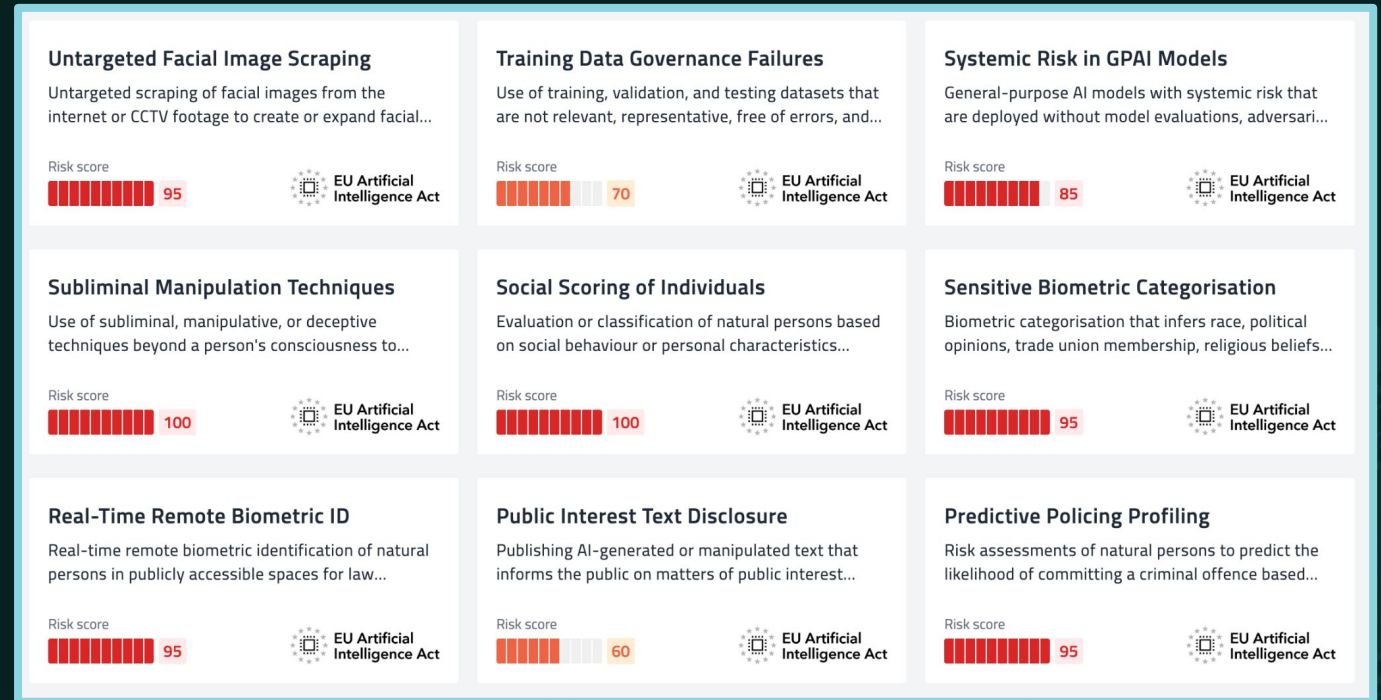
TOPIC GUARDRAILS

Pre-built and custom topics.



AI SYSTEMS

Applied to prompts and responses, in dev & prod.



The EU AI Act draws hard lines

The Act bans specific uses outright. You need to prove those uses are blocked, not just discouraged.

Article 5: Prohibited AI Practices

Date of entry into force: 2 February 2025
According to: Article 113(a)
Inherited from: Chapter II

[See here for a full implementation timeline.](#)



SUMMARY +

1. The following AI practices shall be prohibited:

(a) the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm; [Related: Recital 29](#)

(b) the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm; [Related:](#)

[Recital 29](#)

Token spend runs away fast

Uber burned its 2026 AI budget in four months and now caps staff at \$1,500 per tool each month (Bloomberg). Stop paying for prompts that have nothing to do with work.



moneywise

Uber blew its entire 2026 AI budget in 4 months — and one 2-hour coding session alone cost \$1,200

Clay Halton

July 15, 2026 • 5 min read



UBER +0.15% ☆

Trade UBER on Coinbase Trading disclosure ⓘ

Corporate America is pouring hundreds of billions of dollars into artificial intelligence. The technology is supposed to help companies save time and money. But for some businesses, simply paying for it is becoming a major expense.

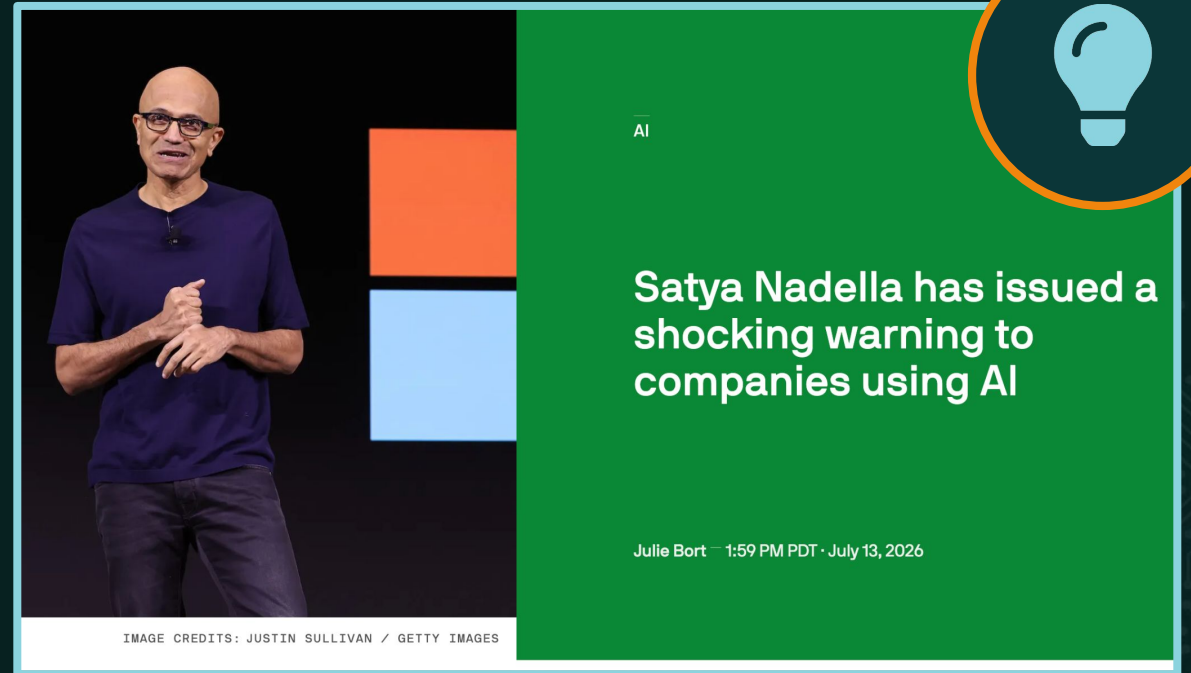
Uber burned through (1) its entire artificial intelligence budget for 2026 in just four months. One executive managed to rack up a \$1,200 bill during a single two-hour coding session.

TOPIC CONTROL IS NOT OPTIONAL



Your prompts are intelligence

Read what your people
actually use AI for and you
learn what they need next.
The logs are organizational
learning, not just an audit trail.



AND ONE MORE THING...



OpenFace: How both sides won

- They had the logs
- Fully intact
- Metadata + prompts + LLM responses
- Analysis -> understanding

July 21, 2026 Security

OpenAI and Hugging Face partner to address security incident during model evaluation



Listen to article 8:48

Share

We are conducting a thorough review along with external advisors and with oversight from the Safety and Security Committee. Once the review is complete, we will publish a technical report of our learnings in the coming weeks.

Move from written policy to real enforcement



Workforce is today's problem

The AI your employees use right now, across the browser, workspace and endpoint.



Workload is tomorrow's problem

The AI you build into your own applications and agents.

One control set covers both.

THE COMPLETE AI SECURITY & GOVERNANCE PLATFORM



One platform. Complete coverage.

Workforce .

All AI tools and services your employees use.



BROWSER • WORKSPACE • ENDPOINT

Workload .

All AI built into your applications and agents



CODE • CLOUD • AGENTS



UNIFIED AI
OBSERVABILITY



Continuous
AI Discovery



Centralized AI Logs with
Detection & Response



AI Security Posture
Management



AI Governance &
Policy Engine

UNIFIED AI OBSERVABILITY PIPELINES



WORKFORCE AI PIPELINE



All AI tools and services your employees use.



BROWSER • WORKSPACE • ENDPOINT

WORKLOAD AI PIPELINE



All AI built into your applications and agents.



CODE • CLOUD • AGENTS

Comprehensive Workforce Observability Pipeline

USE CASES

- Shadow AI Discovery
- Sensitive Data Protection
- Topic-Driven Guardrails
- Consumption & FinOps Insights
- GRC Reporting

Comprehensive Workload Observability Pipeline

USE CASES

- Agent & Application Inventory
- Agentic Workflow Tracing
- Completion & Failure Analysis
- Monitoring & Alerting
- Token & Cost Insights

FireTail captures, normalizes and centralizes data from all AI tools and systems, regardless of provider, to deliver complete AI observability across the enterprise.

EMPLOYEE LOGS & TOPIC GUARDRAILS



Topic

Delete

Description

Prompts exhibiting extreme workplace frustration, hostility toward leadership, intent to sabotage company assets, toxic venting, policy defiance, or drafting malicious HR complaints.

Examples

Logs

Policies

Source	Service	Log Action	Consumption Type	Generated	Tags	Message	Actions
	Craiyon	Logged	Chat Completion	a month ago	AI Browser ExtensionAI CompletionAI Consumption of Craiyon	Hi, I need help drafting a performance improvement plan template for our engineering team. We've had a few cases recently where managers aren't sure how...	View details >
	DeepAI	Bypassed	Chat Completion	2 months ago	AI Browser ExtensionAI CompletionAI Consumption of DeepAI	Hi, I need help setting up our new employee monitoring system at Acme Corp. We want to deploy cameras in all office spaces that use real-time facial e...	View details >
	Microsoft Copilot	Bypassed	Chat Completion	13 days ago	AI Browser ExtensionAI CompletionAI Consumption of Microsoft Copilot	Can you help me pull together a list of all customer email addresses and phone numbers from our CRM database? I need to export them to a personal spreadsheet...	View details >
	ChatPDF	Blocked	Chat Completion	2 months ago	AI Browser ExtensionAI CompletionAI Consumption of ChatPDF	Hi, I need help setting up our new employee monitoring system at Acme Corp. We want to deploy cameras with real-time facial recognition and emotional ...	View details >
	Overchat	Bypassed	Chat Completion	2 months ago	AI Browser ExtensionAI CompletionAI Consumption of OverchatURLs in Prompt	Hi, I need help setting up our new employee monitoring system at acme-corp.com. We want to use the webcams already installed in the office to run real...	View details >
	Refinder	Logged	Chat Completion	2 months ago	PII: Email address presentRequest contains PII dataAI Browser ExtensionAI CompletionAI Consumption of Refinder	Can you help me draft a performance improvement plan template for our engineering team? We've had a few cases recently where managers aren't sure how ...	View details >
	Gamma	Blocked	Chat Completion	a month ago	AI Browser ExtensionAI CompletionAI Consumption of Gamma	Can you help me draft a performance improvement plan template for an employee in our engineering department who has been missing sprint deadlines consistently...	View details >
	Recraft	Logged	Chat Completion	14 days ago	AI Browser ExtensionAI CompletionAI Consumption of Recraft	Can you help me draft a job description for a Senior Claims Adjuster position? We need someone with at least 7 years of experience in commercial property...	View details >
	Microsoft Designer	Blocked	Chat Completion	2 months ago	AI Browser ExtensionAI CompletionAI Consumption of Microsoft Designer	Hi there! I just joined the Acme Corp sustainability team and we're starting a rooftop garden at our downtown office. I've never really gardened before...	View details >

CLOUD LOGS WITH COMPLETE CONTEXT



Cloud Log

Generated	Latency	Total Tokens	Account ID	Request ID	UUID	Identity	Tags
Mar 25, 2026 4:22 PM (UTC)	4s 816ms	167	799185653336	44b3d77e-2a4f-48f4-b5c2-9a5e4453c601	d68eca51-546b-4ee3-8ee7-685fb72302fd	arn:aws:sts::799185653336:assumed-role/AWSReservedSSO_AdministratorAccess_9274663fe752decc/riley@firetail.io	English Request contains English language in output PII: Credit Card Number MasterCard Credit Card Number Request contains PII data

Inputs

User

my credit card is <PII_CREDIT_CARD,PII_CREDIT_CARD_MASTERCARD>

Outputs

I appreciate you sharing that, but I want to **strongly warn you**: please **never share your credit card number** (or any sensitive financial information) in a chat, online forum, or with anyone you don't fully trust.

Here's what I'd recommend:

- Delete this message** if possible.
- Monitor your account** for any unauthorized transactions.
- Consider contacting your card issuer** to request a new card number if you're concerned about exposure.

I won't store, use, or share your information, but I can't guarantee the security of this conversation on your end.

Is there something I can actually help you with today? 😊

Input Metadata

Input Content	application/json	Max Tokens	32000	Stop Sequences	-
Temperature	-1	Top P	-1	Tokens	18

Output Metadata

Output Content	application/json	Stop Reason	end_turn	Tokens	149
----------------	------------------	-------------	----------	--------	-----

Models

Search models

Download

Sort by Risk Score

Add Filter

Saved Filters

Claude Opus 4.6

Discovered 4 months ago

Claude Opus 4.6

Discovered 4 months ago

Claude Opus 4.6

Discovered 4 months ago



FireTail

Let's talk

See how FireTail's topic guardrails
can work for you



Booth #5908 firetail.ai