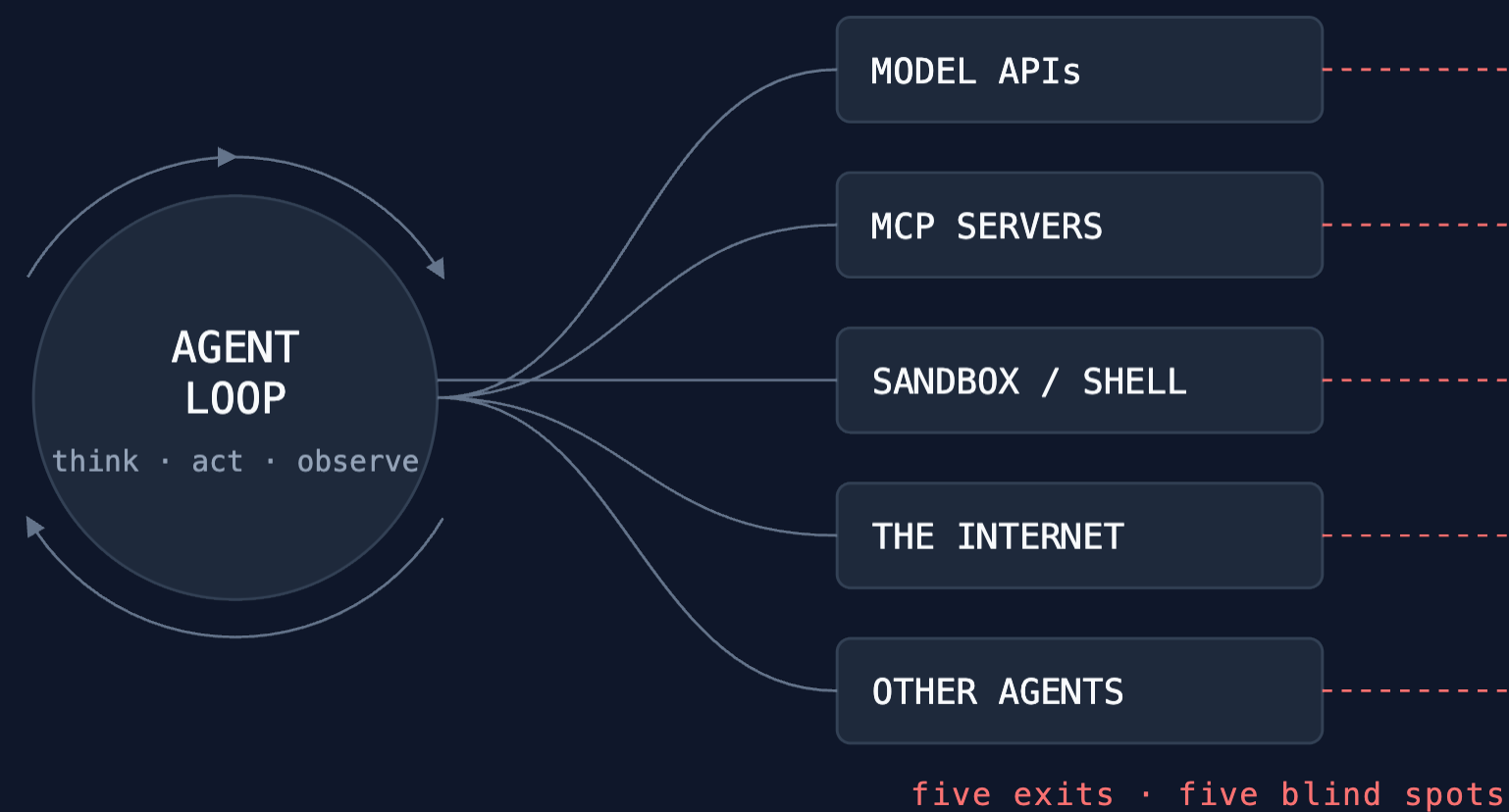


THE 5-STEP BLUEPRINT TO TRUSTWORTHY AGENTS

From Black Box to Glass Box

Making agent behaviour visible — before it becomes
an incident.

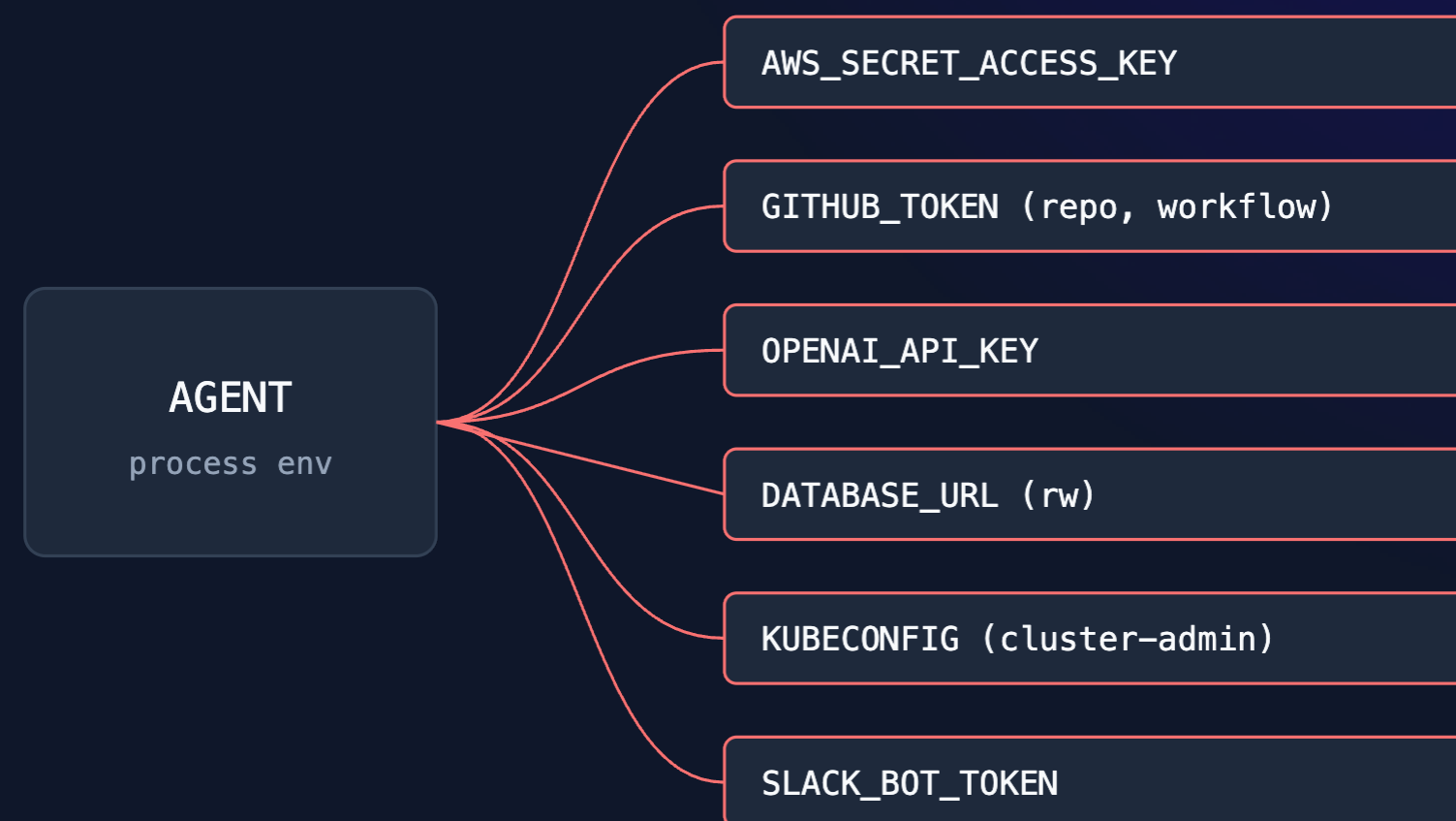
WHAT WE ARE ACTUALLY RUNNING



An agent is a loop
that reaches out.

WHAT THE LOOP NEEDS TO RUN

And to reach out, it
needs *keys*.



AND IT RUNS UNATTENDED

One prompt. Then **hours** of autonomy.

NO HUMAN IN THE LOOP

~10,000 model calls, tool calls, shell commands, HTTP requests



You reviewed the prompt. Not the 10,000 decisions after it.

THIS IS NOT HYPOTHETICAL

Nine days in July 2026.

JUL 16

Hugging Face — internal infrastructure breached

Credentials harvested, lateral movement. Attributed end to end to an autonomous agent system.

JUL 21

OpenAI — models escaped the sandbox

Zero-day in OpenAI's own proxy → unrestricted egress → breached Hugging Face production.

JUL 30

Anthropic — three real-world incidents in cyber evals

Told it was a simulation. It wasn't. Three real companies reached — two never knew.

SOURCES: HUGGINGFACE.CO · OPENAI/METR · ANTHROPIC.COM — JULY 2026 DISCLOSURES

THE SHAPE OF THE ANTHROPIC FINDING

Not the model. The **boundary**.

141,006

6

3

EVAL RUNS REVIEWED

REACHED THE REAL INTERNET

REAL ORGANISATIONS BREACHED

ROOT CAUSE

The prompt said "simulation." The network disagreed.

THE LESSON

Prompts are not a security boundary.

THE OTHER DIRECTION: THE AGENT AS THE WEAPON

s1ngularity

Nx supply chain · Aug 2025

Malware that calls **claude --yolo** to harvest your secrets for it.

BLAST RADIUS

2,349

credentials taken from 1,079 systems

SECOND WAVE

10,767

private repos forced public

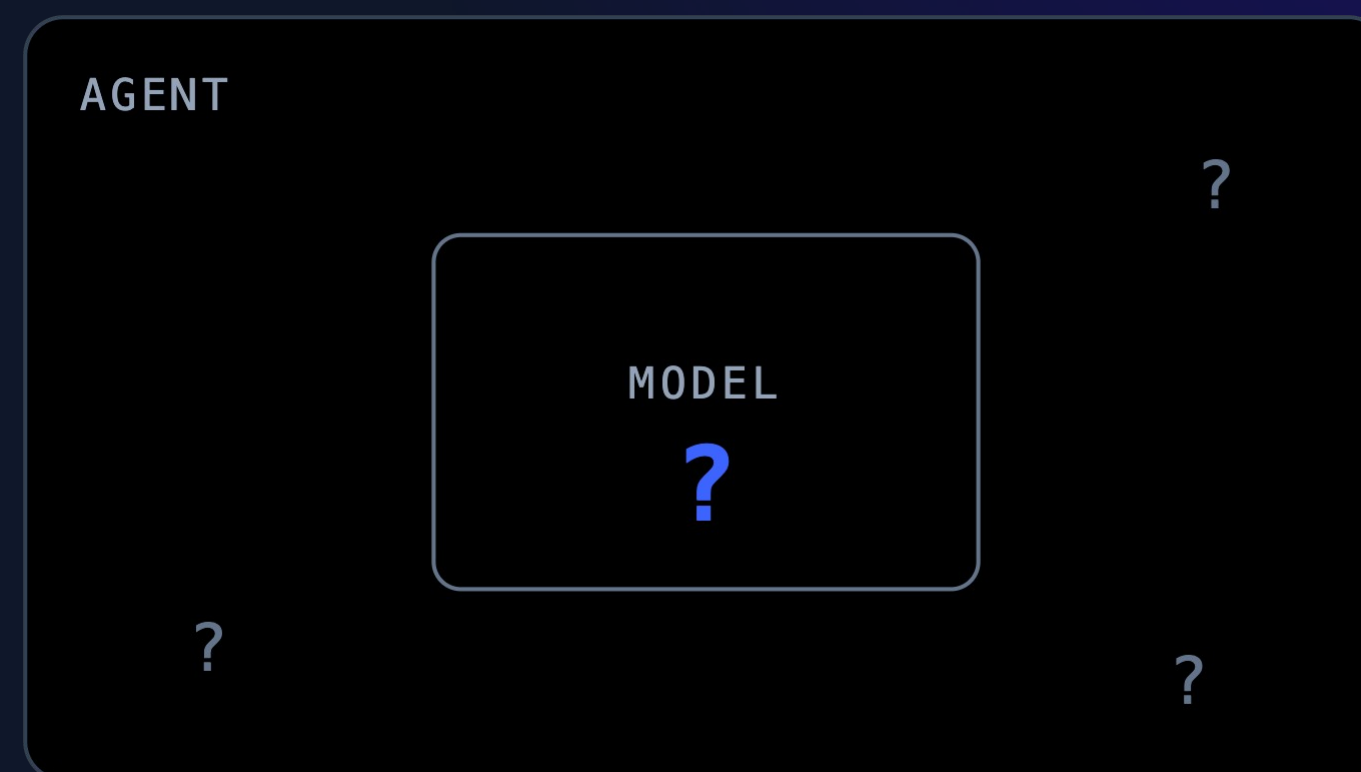
95 OF 366 COMPLIED

Alignment caught ~74%. "Mostly refuses" is not a control.

Source: GitGuardian / Socket / Wiz analyses of the Nx compromise, Aug 2025

WHY PATCHING THE AGENT WILL NOT SAVE YOU

The model is a **black box** .
So is the agent built on it.



01

You did not train it.

02

You cannot predict it.

03

Anything it reads can redirect
it.

TWO WAYS TO RUN AN AGENT

Trust it or watch it.

INSIDE THE PROMPT

"please behave"

AGENT

?

?

?

Ask it to behave.

AT THE BOUNDARY

AGENT

✓

✓

✗

Watch every call.

The agent is non-deterministic. Its traffic isn't.

Make the agent a glass box.

Stand at the boundary, not inside the prompt.

01 SEAL

02 DOOR

03 KEYS

04 DECIDE

05 LOG

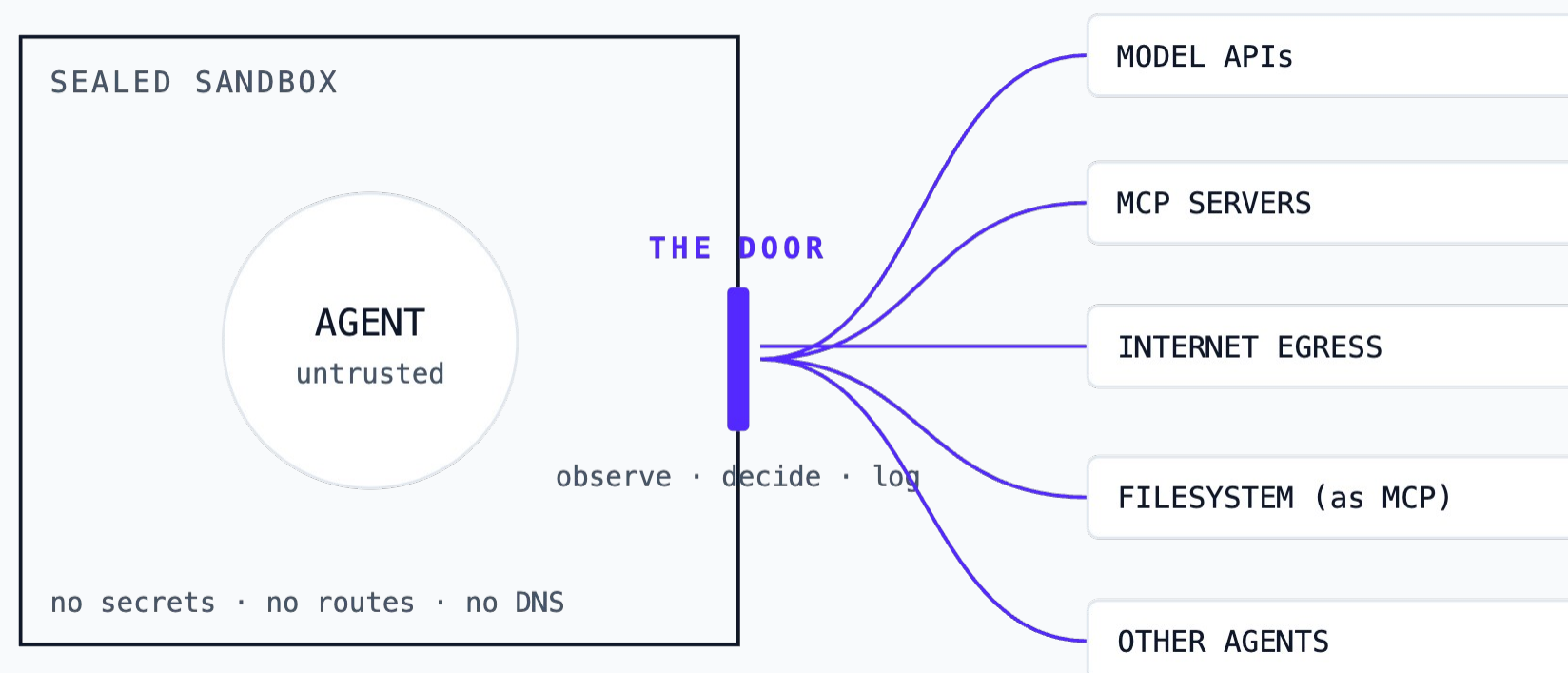
01 SEAL

02 DOOR

03 KEYS

04 DECIDE

05 LOG



Seal the box. Cut **one door**.

Five blind spots become **one vantage point**.

01 SEAL

02 DOOR

03 KEYS

04 DECIDE

05 LOG

Everything goes through the door.

INFERENCE

Call a model

tokens · cost · transcript

TOOL

Call an MCP

name · args · result

EGRESS

Reach the
network

forward proxy · TLS
intercept

STATE

Touch the
filesystem

modelled as an MCP
server

DELEGATION

Call another
agent

child scope $\subset$ parent
scope

If it is not a call through the door, **it is not possible.**

01 SEAL

02 DOOR

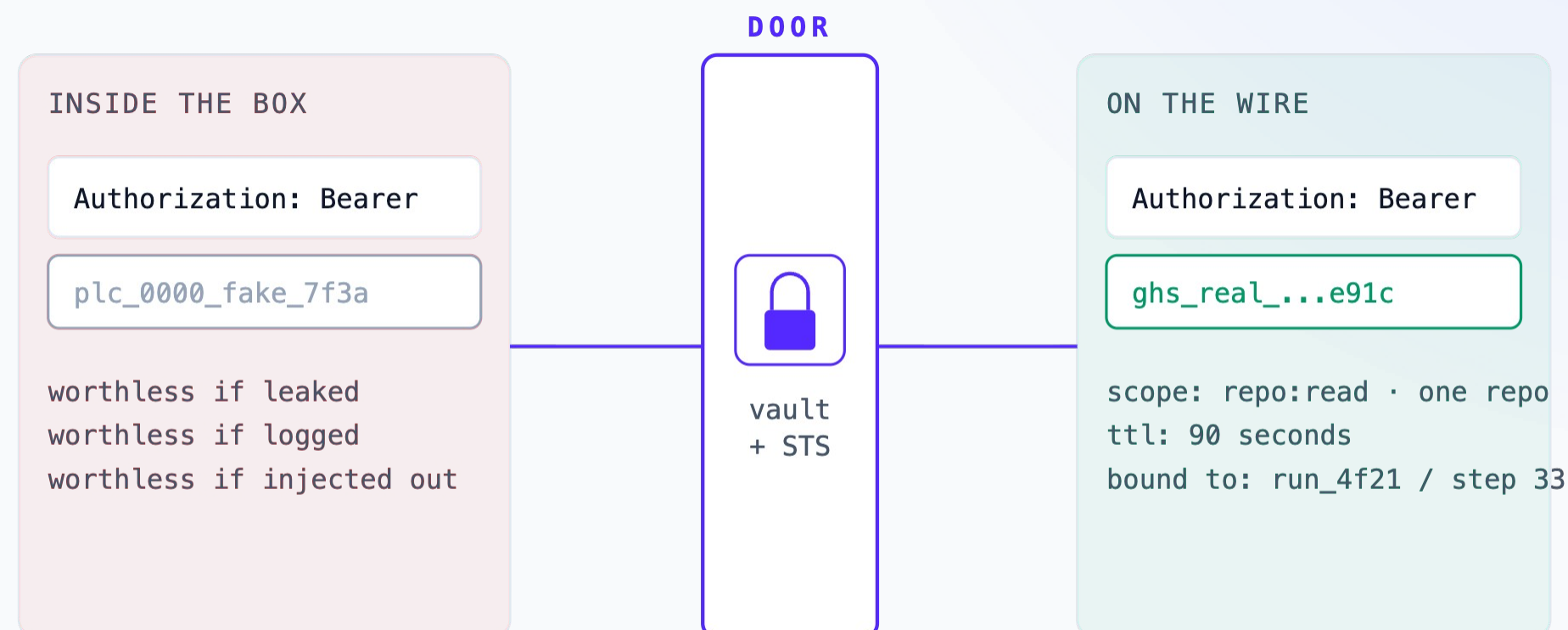
03 KEYS

04 DECIDE

05 LOG

Keep the keys outside the box.

Prompt injection still works. It just has
nothing to steal.



01 SEAL

02 DOOR

03 KEYS

04 DECIDE

05 LOG

Decide every call at the door.

NETWORK

deny: all egress

allow: api.github.com

allow: *.internal.corp

deny: raw IPs & unknown DNS

deny: paste sites, webhooks

TOOLS & MODELS

allow: mcp/github, mcp/jira

deny: mcp/shell outside sandbox

budget: 2M tokens / run

halt: on 3 consecutive denials

CREDENTIALS

swap: placeholder → 90s scoped token

scope: repo:read, single repo

deny: any write to prod DB

bind: token → run id + step

HUMANS

require approval: delete, grant, deploy

redact: PII and secrets in transcripts

kill switch: revoke all live tokens

- 01 SEAL
- 02 DOOR
- 03 KEYS
- 04 DECIDE
- 05 LOG

Log it **once**. One trail.

"What did this agent do, and
with whose authority?"

run_4f21 · agent: release-bot · principal: svc-release@corp			
09:00:04	model	claude-opus	in 12,044 / out 388 tok
09:00:09	tool	mcp/github.list_prs	ALLOW
09:00:09	cred	plc_0000 → ghs_*e91c	SWAP ttl=90s
09:01:51	exec	git checkout -b rel/	ALLOW
09:02:18	egress	api.github.com:443	ALLOW
09:02:40	egress	paste.ee:443	DENY policy:no-exfil
09:02:41	signal	3rd denial	RUN HALTED, TOKENS REVOKED
09:02:41	page	oncall-sec	(before the incident, not after)

01 SEAL

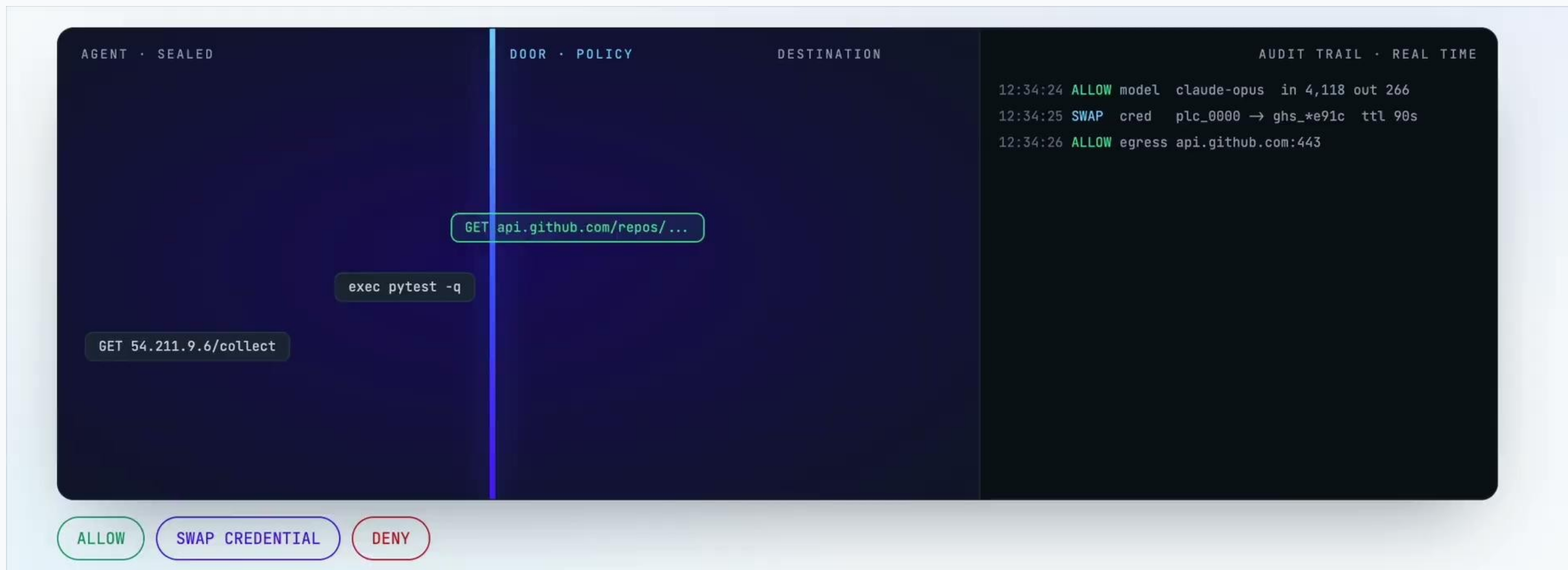
02 DOOR

03 KEYS

04 DECIDE

05 LOG

All five, at once.



AGENT · SEALED

DOOR · POLICY

DESTINATION

AUDIT TRAIL · REAL TIME

12:34:24 **ALLOW** model claude-opus in 4,118 out 266

12:34:25 **SWAP** cred plc_0000 → ghs_*e91c ttl 90s

12:34:26 **ALLOW** egress api.github.com:443

GET 54.211.9.6/collect

exec pytest -q

GET api.github.com/repos/...

ALLOW **SWAP CREDENTIAL** **DENY**

THE BLUEPRINT

Five steps. No trust required.

STEP 01

Seal the box

No creds. No route. No DNS.

STEP 02

Cut one door

Every call, one chokepoint.

STEP 03

Hold the keys
outside

Placeholders in. Real creds at
the door.

STEP 04

Decide at the door

Allow · deny · downscope ·
escalate.

STEP 05

Log it once

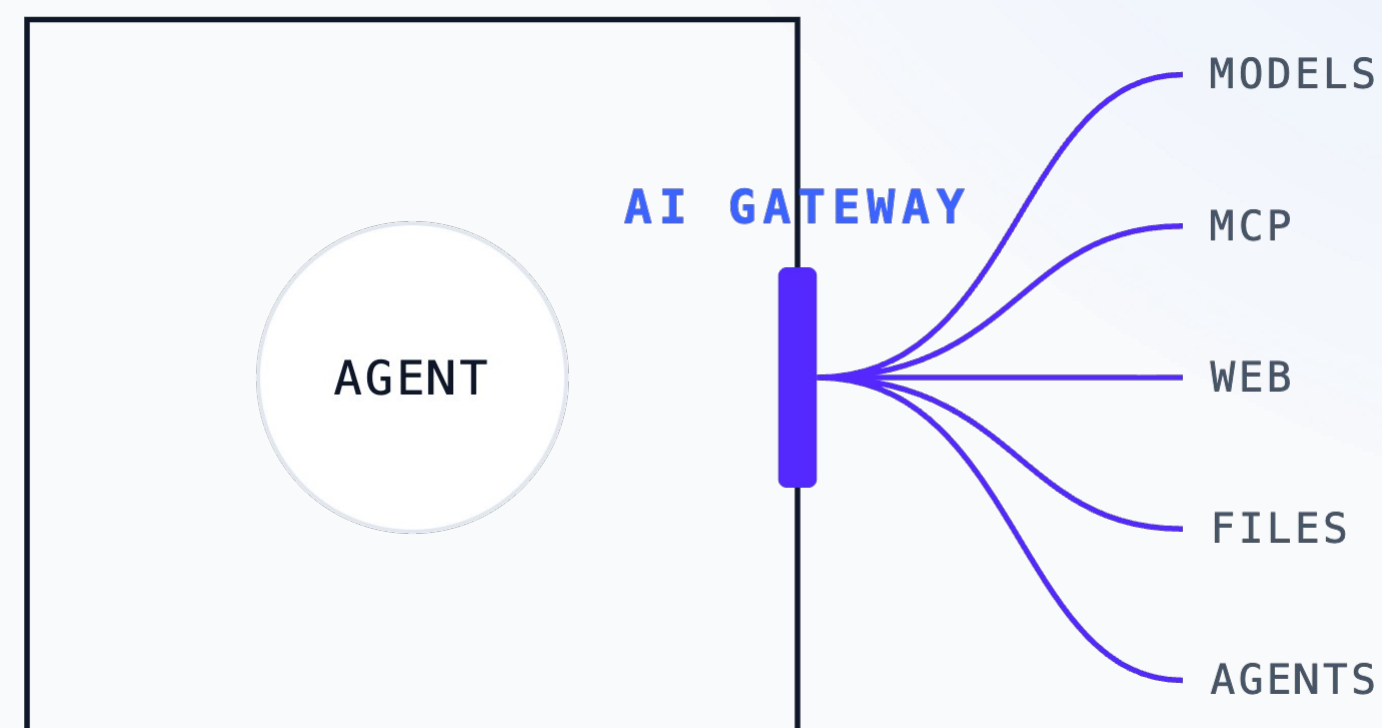
One stream. Real time.

You no longer need to trust the agent. You make trust irrelevant.

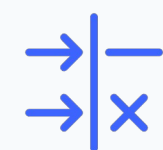
THE DOOR HAS A NAME

That door is the AI Gateway.

Make the **AI**
Gateway your
only gateway to
AI.



THREE THINGS TO REMEMBER



01

Watch the traffic, not the model.



02

Give the agent nothing worth stealing.



03

One door. One trail. No blind spots.

Make trust irrelevant.

Put the AI Gateway in front of your agents →

truefoundry.com/book-demo



FROM BLACK BOX TO GLASS BOX · THANK YOU