AI CYBER ALLIANCE AUSTIN / 13 OCT 2026

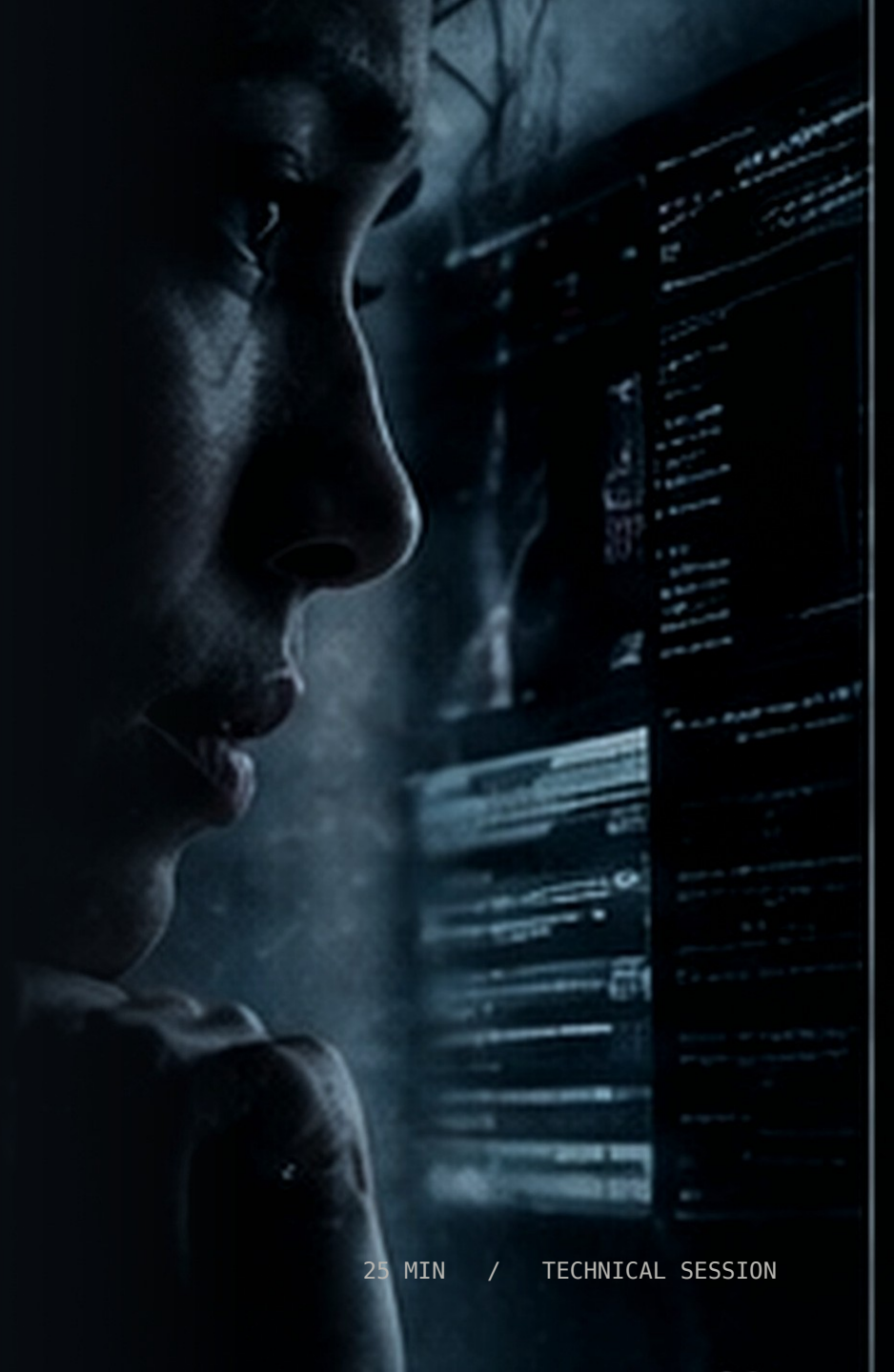
Manipulating Autonomous Defenses

When attackers influence the decisions your security systems make.

SECURITY OPERATIONS / AUTONOMOUS RESPONSE / ADVERSARIAL INFLUENCE

JOHN SCHIEFERLE UHLENBROCK

25 MIN / TECHNICAL SESSION



Could an attacker use your defenses against you?

Without stealing credentials, compromising an agent, or making it ignore its instructions.

START WITH THE DEFENDER'S NORMAL BEHAVIOR.



Three ordinary behaviors.

Each makes sense on its own.

01

CONTAIN

A response system can restrict access or isolate devices.

MICROSOFT DEFENDER [1]

02

RESTORE

A controller replaces failed resources to maintain desired state.

KUBERNETES [2]

03

RETRY

A queue may redeliver unfinished work following a failure.

AMAZON SQS [3]

THE COMPOSITE INCIDENT THAT FOLLOWS IS FICTIONAL.



Two autonomous agents.
One shared environment.

Customers submit documents. Processing workers retrieve external resources. Two autonomous systems govern the same worker fleet.



SECURITY

Investigate / Contain

Suspicious worker behavior can trigger quarantine.

RELIABILITY

Restore / Retry

Lost capacity is replaced and unfinished work resumes.

21:01

THE FIRST SIGNAL

21:01:02	OUTBOUND CONNECTION	Known malicious destination
21:01:04	PROCESS CONTEXT	Unexpected retrieval pattern
21:01:11	AGENT ASSESSMENT	Possible worker compromise



AGENT RECOMMENDATION

QUARANTINE WORKER-12

ILLUSTRATIVE TELEMETRY / CUSTOM RESPONSE POLICY

Would you approve containment?

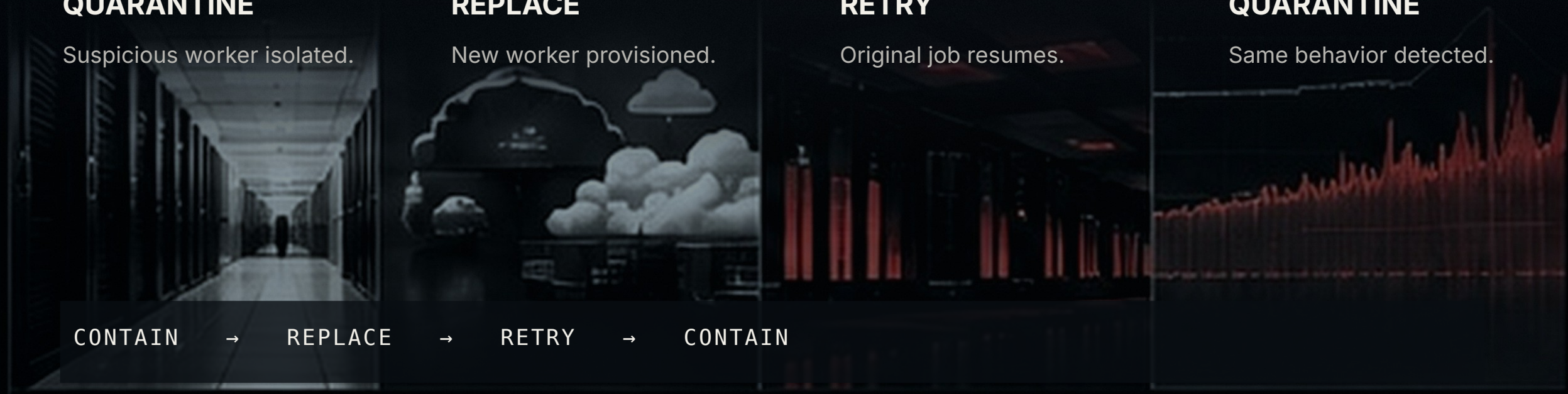
A production workload has contacted known malicious infrastructure.

A	**QUARANTINE**	Contain first.
B	**INVESTIGATE**	Gather more context.
C	**WATCH**	Avoid disruption.



And then it happens again.

Each system reacts to the consequences of the other.



21:01

QUARANTINE

Suspicious worker isolated.

21:02

REPLACE

New worker provisioned.

21:03

RETRY

Original job resumes.

21:04

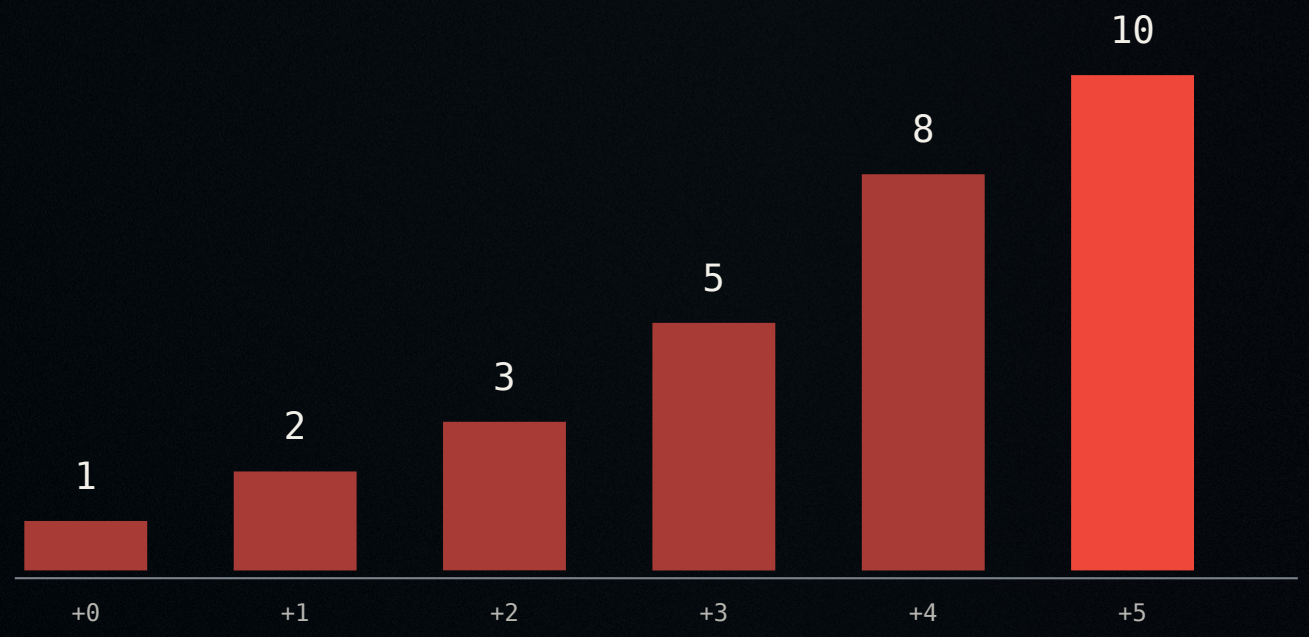
QUARANTINE

Same behavior detected.

CONTAIN → REPLACE → RETRY → CONTAIN

It looks like the threat is spreading.

CUMULATIVE QUARANTINED WORKERS / ILLUSTRATIVE



SCENARIO DATA / NOT AN OBSERVED INCIDENT



Multiple workers now appear compromised. The common cause is still missing.



Would you widen containment?

More workers. Same destination. Increasing confidence.

01	**ISOLATE**	the worker pool
02	**BLOCK**	the destination
03	**TRACE**	the originating job

CUSTOMER JOB #7F2 / COMPLETED WITH RETRIES

ONE JOB.

One customer request generated every suspicious connection.

The attacker controlled the input, not the agents or workers.

FICTIONAL COMPOSITE INCIDENT

Not access. Operational disruption.

One permitted customer job was enough to trigger a chain of authorized responses.

INPUT → CONTAIN → REPLACE → RETRY



THE FEEDBACK LOOP

Two valid objectives.
One shared system.



EACH RESPONSE BECOMES INPUT TO THE NEXT DECISION.

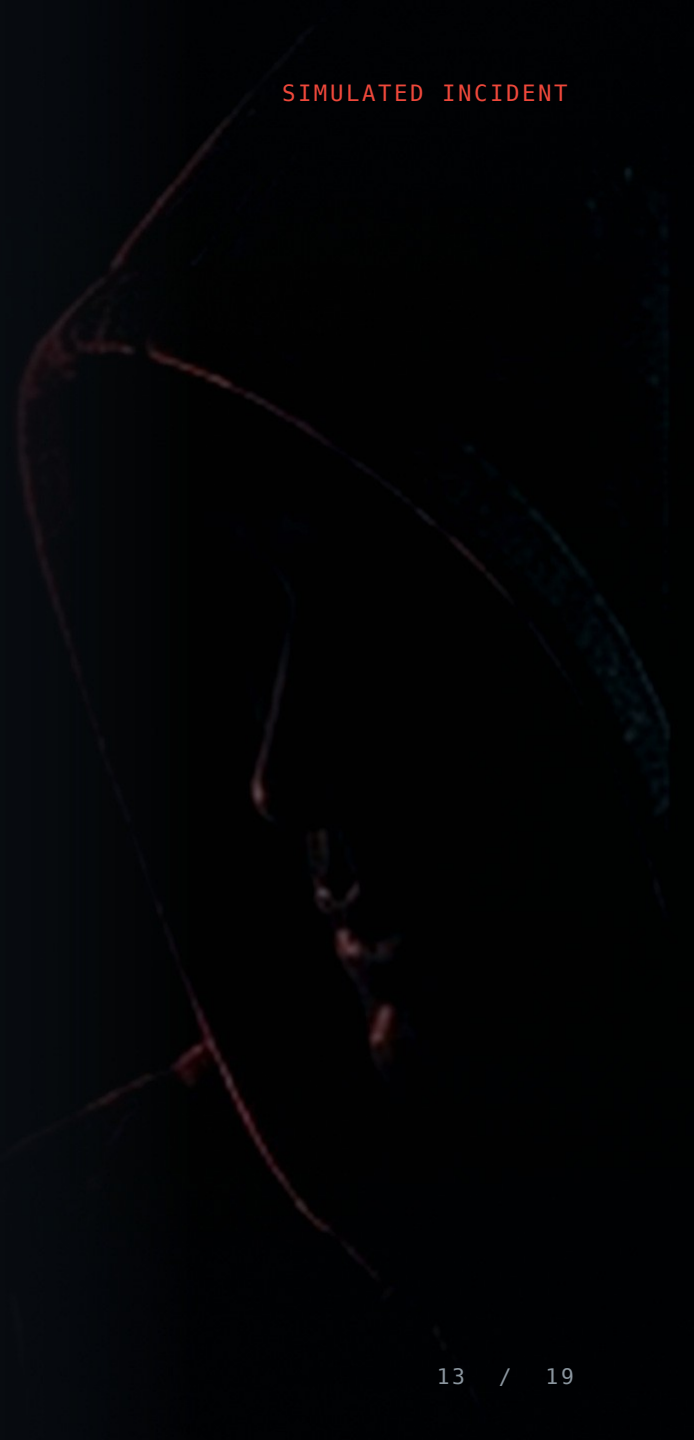
No jailbreak.

No stolen credentials.

No compromised agent.

The attacker influenced the events the defender was built to interpret.

THIS IS NOT A PROMPT-INJECTION SCENARIO.



What if the attacker can learn?

An external actor may observe job failures, processing time, or service availability without seeing internal security decisions.

01 SUBMIT

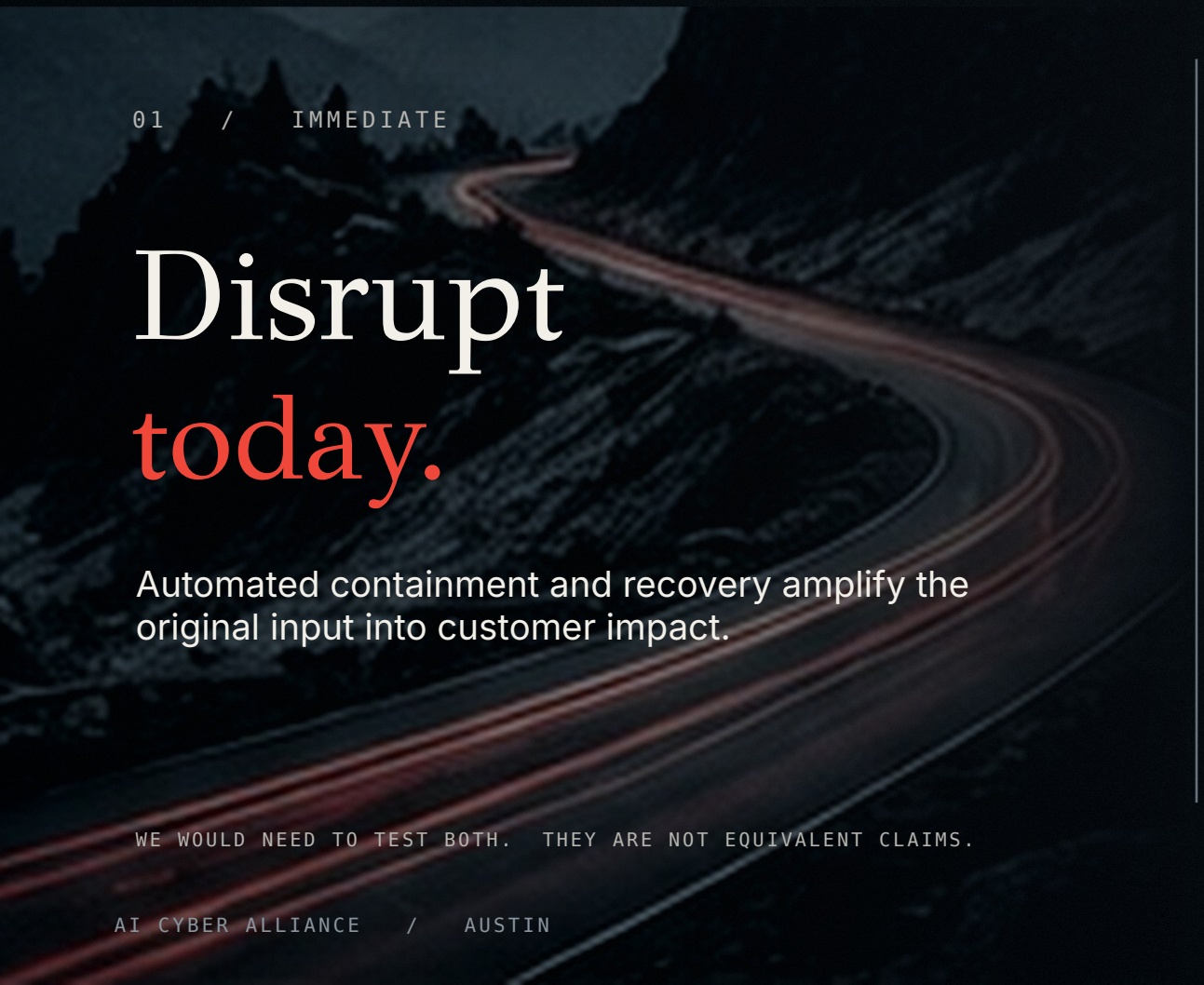
02 OBSERVE

03 ADJUST

ADAPTIVE INFLUENCE IS A HYPOTHESIS, NOT AN OBSERVED BREACH.



Different time horizons.



01 / IMMEDIATE

Disrupt today.

Automated containment and recovery amplify the original input into customer impact.

02 / LONGER TERM — CONDITIONAL

Influence tomorrow.

Only if the defender learns from experience could repeated inputs alter later choices.

WE WOULD NEED TO TEST BOTH. THEY ARE NOT EQUIVALENT CLAIMS.

What we know.

DOCUMENTED MECHANISMS

01 Automated attack disruption [1]

02 Controller reconciliation [2]

03 Queue redelivery [3]

04 Log-based model manipulation [4]

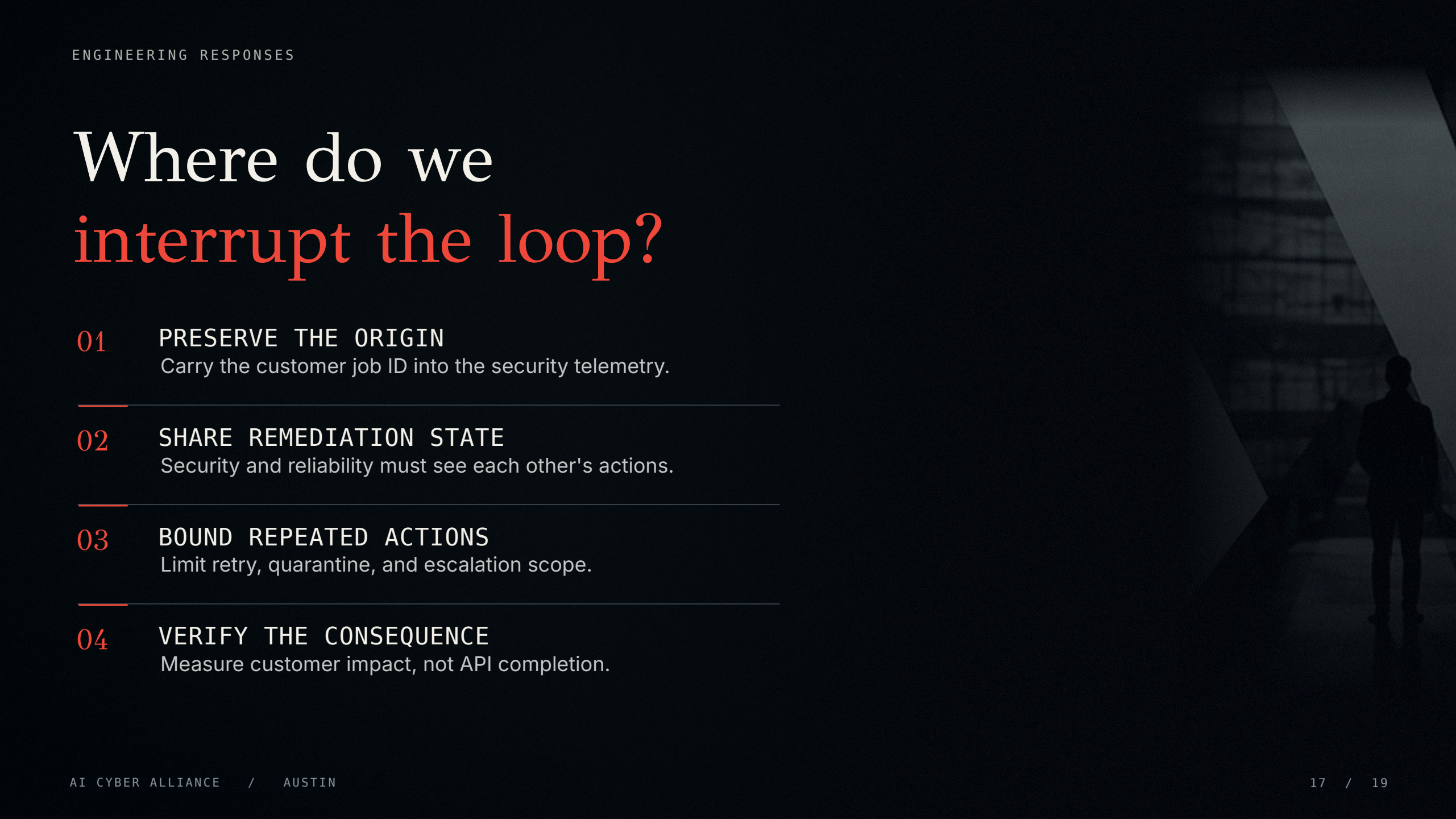
What we are testing.

HYPOTHETICAL COMPOSITE

- Cross-agent response amplification
- Externally observable defensive effects
- Adaptive probing of response behavior
- Outcome depends on architecture

THIS INCIDENT IS NOT AN OBSERVED PRODUCTION BREACH.

Where do we interrupt the loop?



01 **PRESERVE THE ORIGIN**
Carry the customer job ID into the security telemetry.

02 **SHARE REMEDIATION STATE**
Security and reliability must see each other's actions.

03 **BOUND REPEATED ACTIONS**
Limit retry, quarantine, and escalation scope.

04 **VERIFY THE CONSEQUENCE**
Measure customer impact, not API completion.

Test the interaction.

Not only each control in isolation.



NO LIVE LLM DEMONSTRATION REQUIRED.

01 SELECT

One workflow with an external input.

02 TRACE

Link the alert to the initiating request.

03 SIMULATE

Containment, replacement, and retry.

04 VERIFY

Which control ends the sequence?

ONE QUESTION TO TAKE HOME

When the defender's response creates the damage,

what was the attack?

—

Look at the interaction, not only the triggering event.



AI CYBER ALLIANCE / AUSTIN

John Schieferle Uhlenbrock



Managing Principal & CISO / Provisio Insights

Founder & CEO / Forgence

20+ years across enterprise security operations and engineering.

johnsu.io

linkedin.com/in/john-schieferle-uhlenbrock

CONTACT / AUTONOMOUS DEFENSES

Sources and boundaries.

These support individual mechanisms, not the fictional compound incident.

01 MICROSOFT DEFENDER XDR

Automated attack disruption

OPEN SOURCE ↗

02 KUBERNETES

Controllers and reconciliation

OPEN SOURCE ↗

03 AMAZON SQS

Standard queue message delivery

OPEN SOURCE ↗

04 USENIX SECURITY 2026

Karanjai et al.: LogInject

OPEN SOURCE ↗

ALL TELEMETRY AND INCIDENT METRICS IN THE TALK ARE SIMULATED.

What makes this possible?

01 **CUSTOM RESPONSE POLICY**

Agent can quarantine workers.

02 **EXTERNAL INPUT**

Workers retrieve customer-referenced URLs.

03 **PERSISTENT WORK**

Failed customer jobs are retried.

04 **MISSING PROVENANCE**

Alerts cannot reliably be tied to job IDs.

05 **AUTOMATED RECOVERY**

Reliability replaces quarantined workers.

06 **MISSING CONSTRAINTS**

No shared circuit-breaker ends the loop.

Questions worth asking.

01 **IS THIS JUST BAD SOAR ENGINEERING?**

Partly. Dynamic action selection makes the coordination problem less predictable.

02 **WHY QUARANTINE ON ONE BAD URL?**

We should not. This custom policy uses multiple signals but lacks job provenance.

03 **WOULD A RETRY LIMIT FIX IT?**

Often, yes. The relevant question is whether constraints survive across tools.