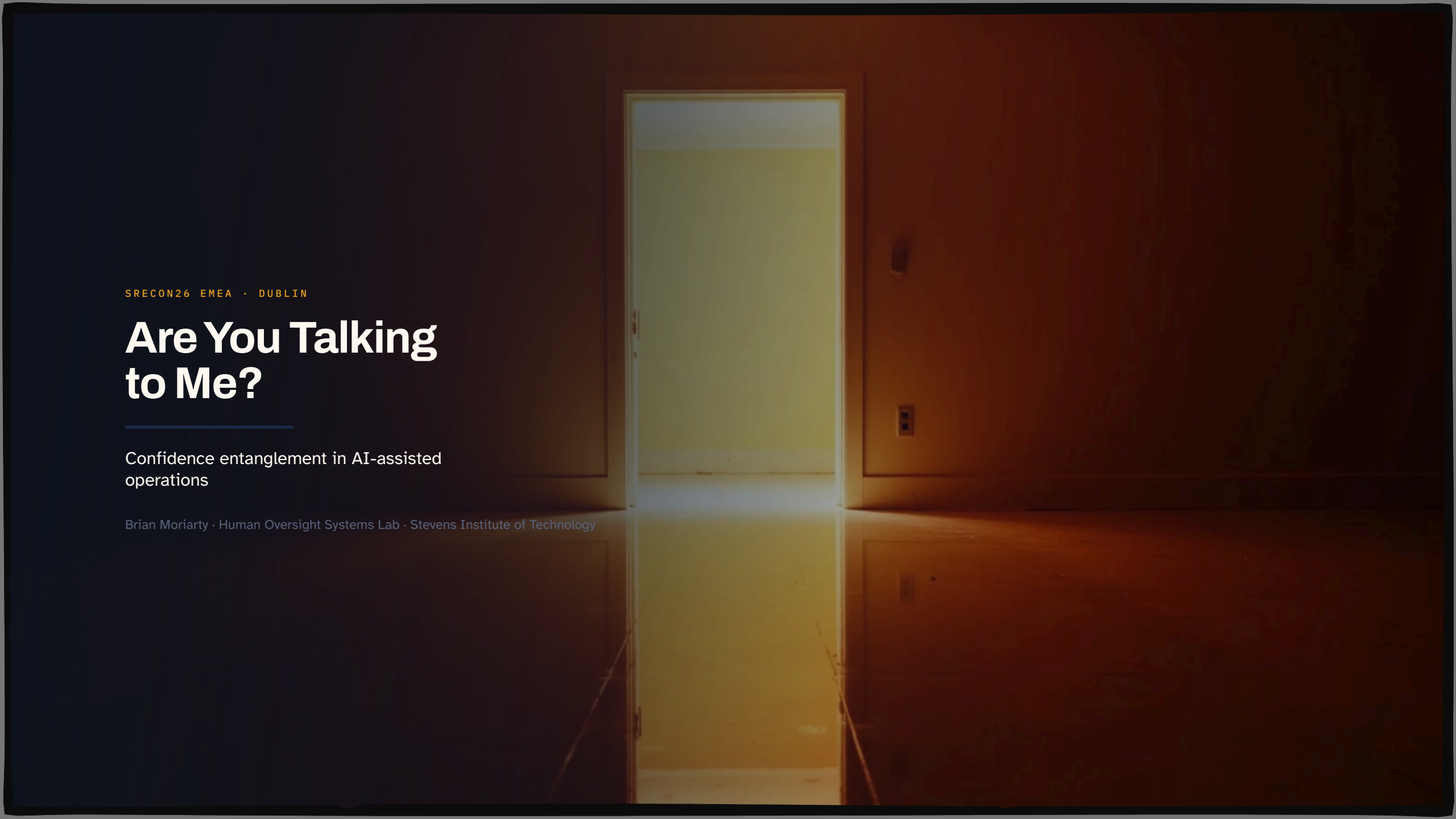
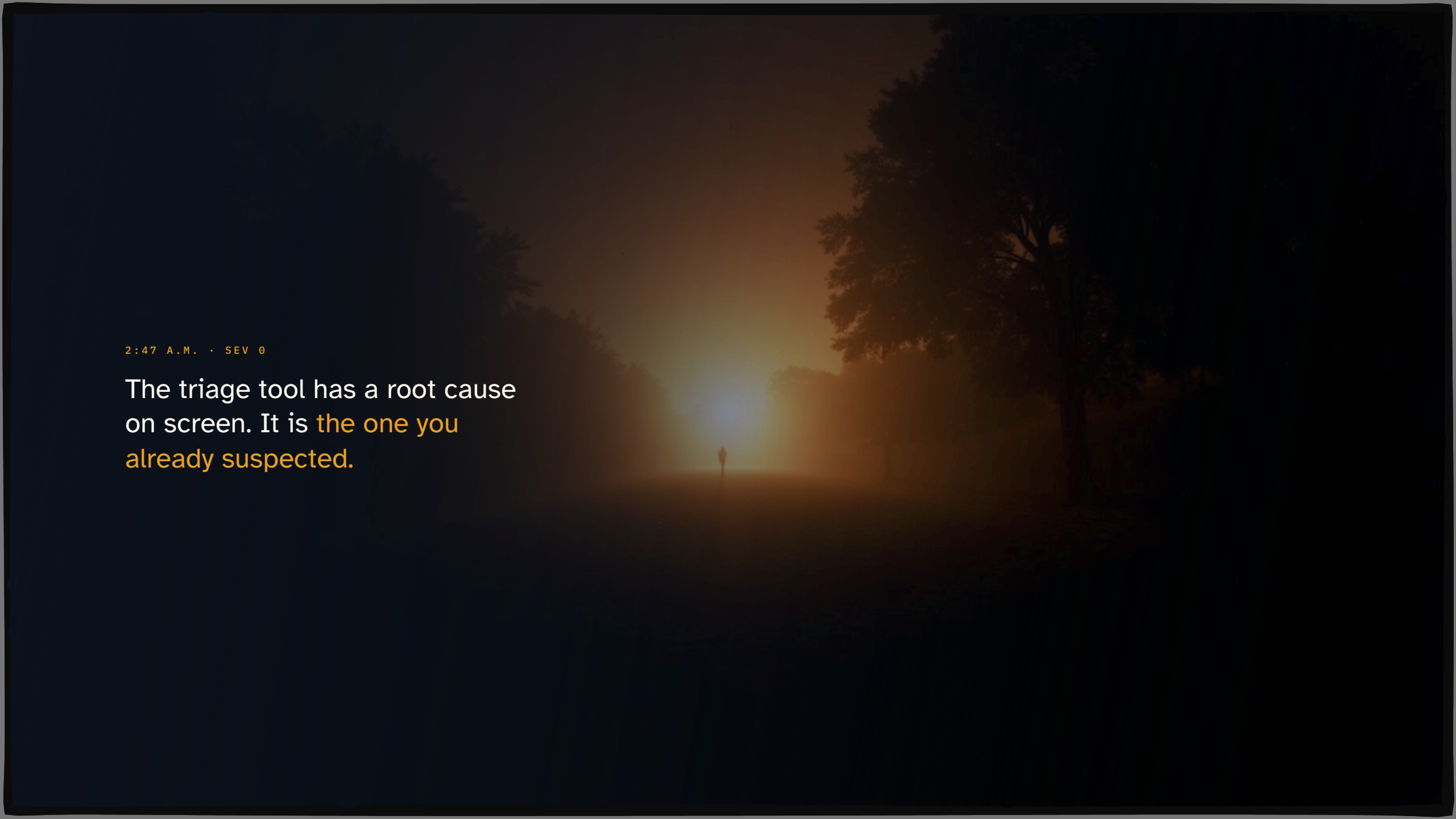
SRECON26 EMEA · DUBLIN

Are You Talking to Me?

Confidence entanglement in AI-assisted operations

Brian Moriarty · Human Oversight Systems Lab · Stevens Institute of Technology

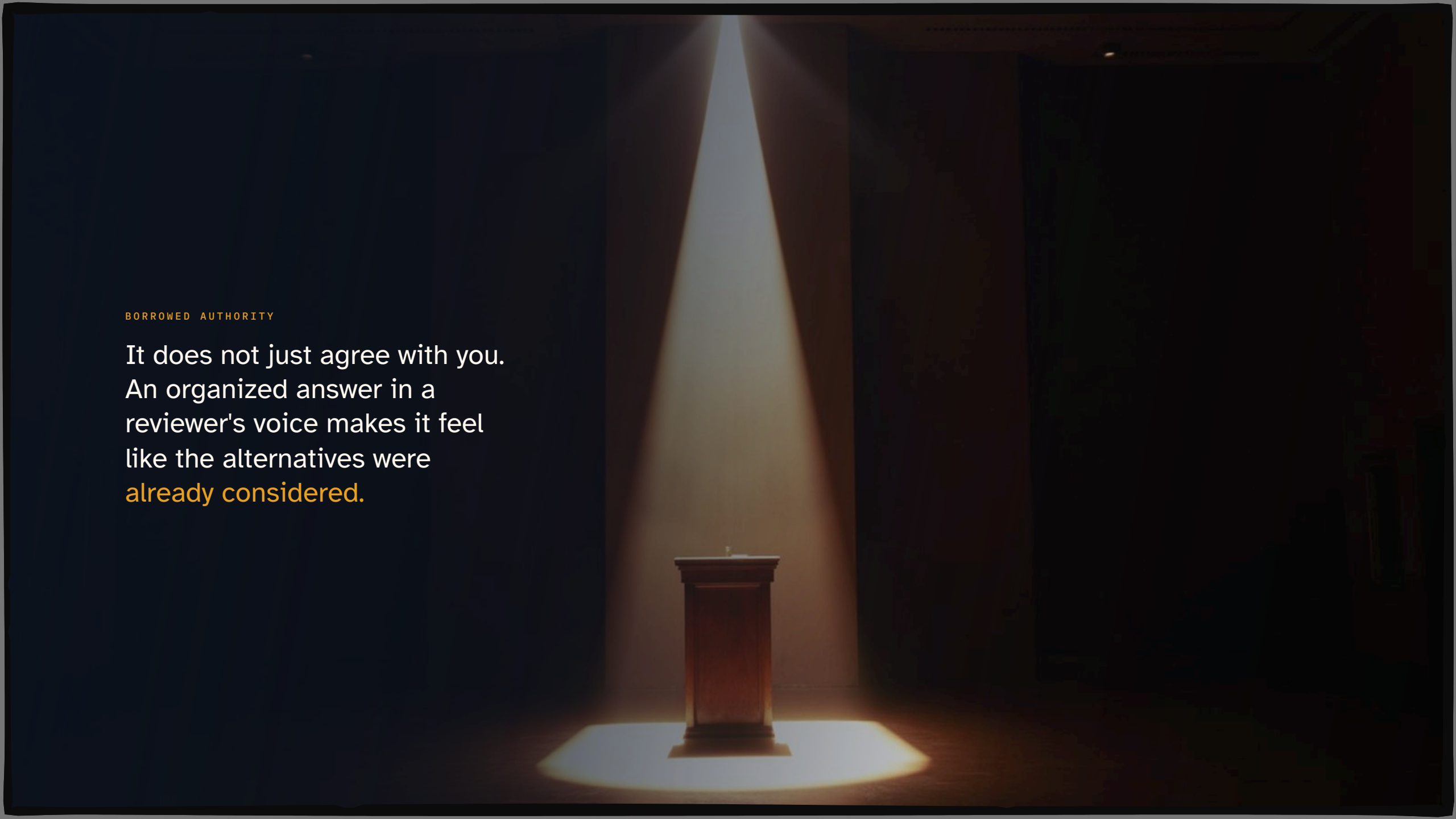



2:47 A.M. · SEV 0
The triage tool has a root cause on screen. It is the one you already suspected.


"You talkin' to me?"
A MIRROR THAT CANNOT PUSH BACK

BORROWED AUTHORITY

It does not just agree with you. An organized answer in a reviewer's voice makes it feel like the alternatives were already considered.

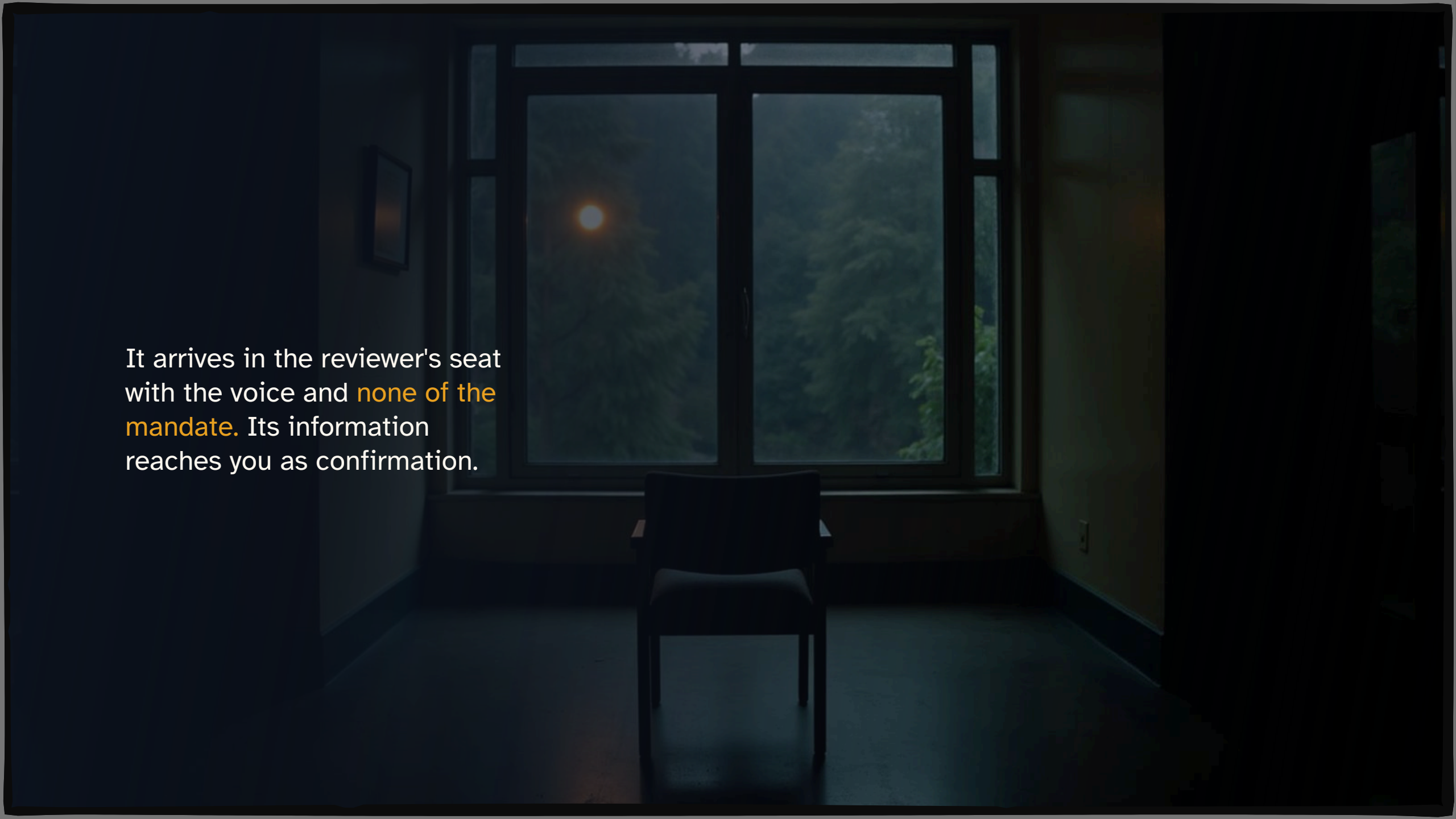


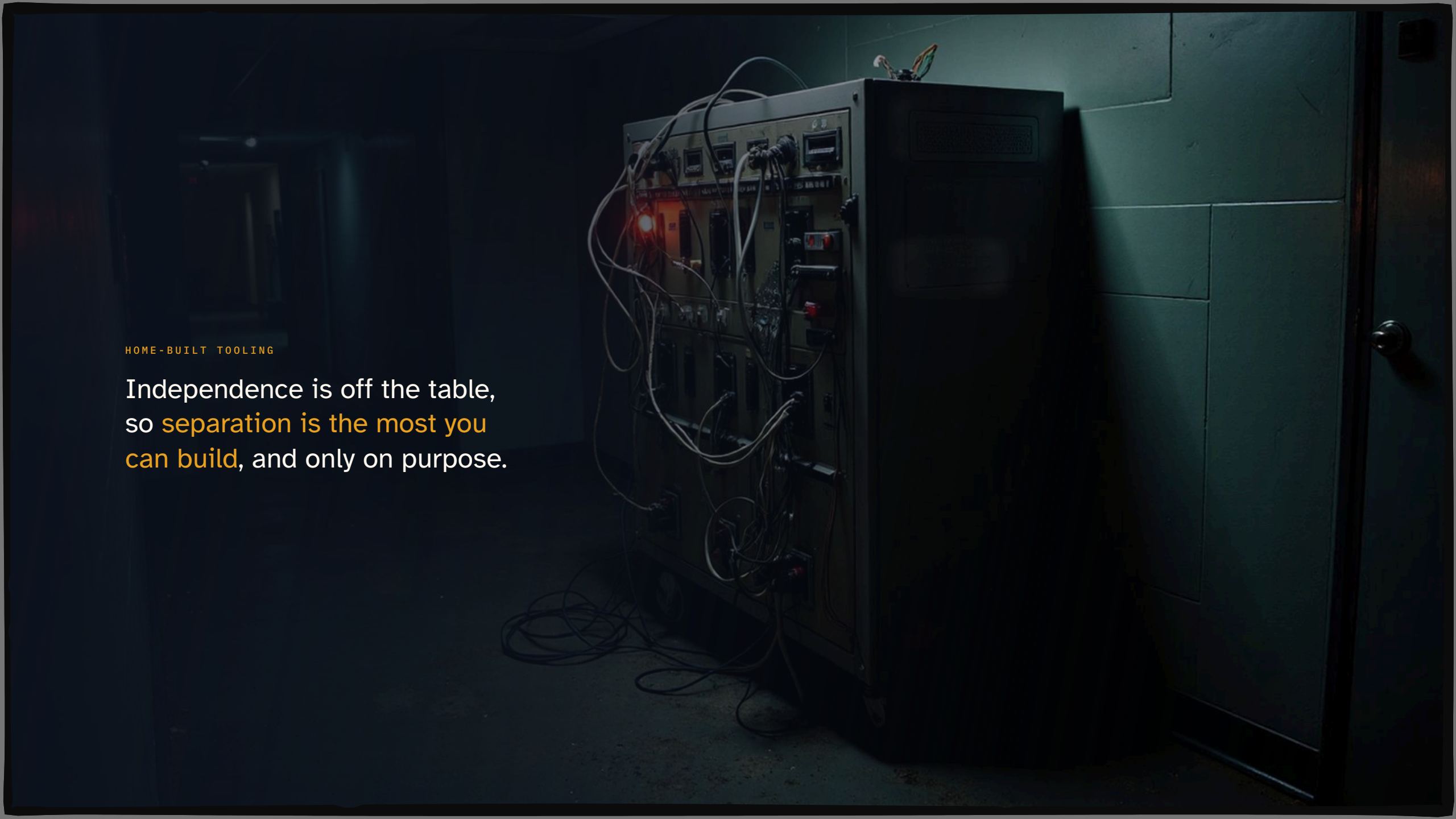


SECOND OPINIONS

You cannot conjure independence at 2am. Build separation instead.

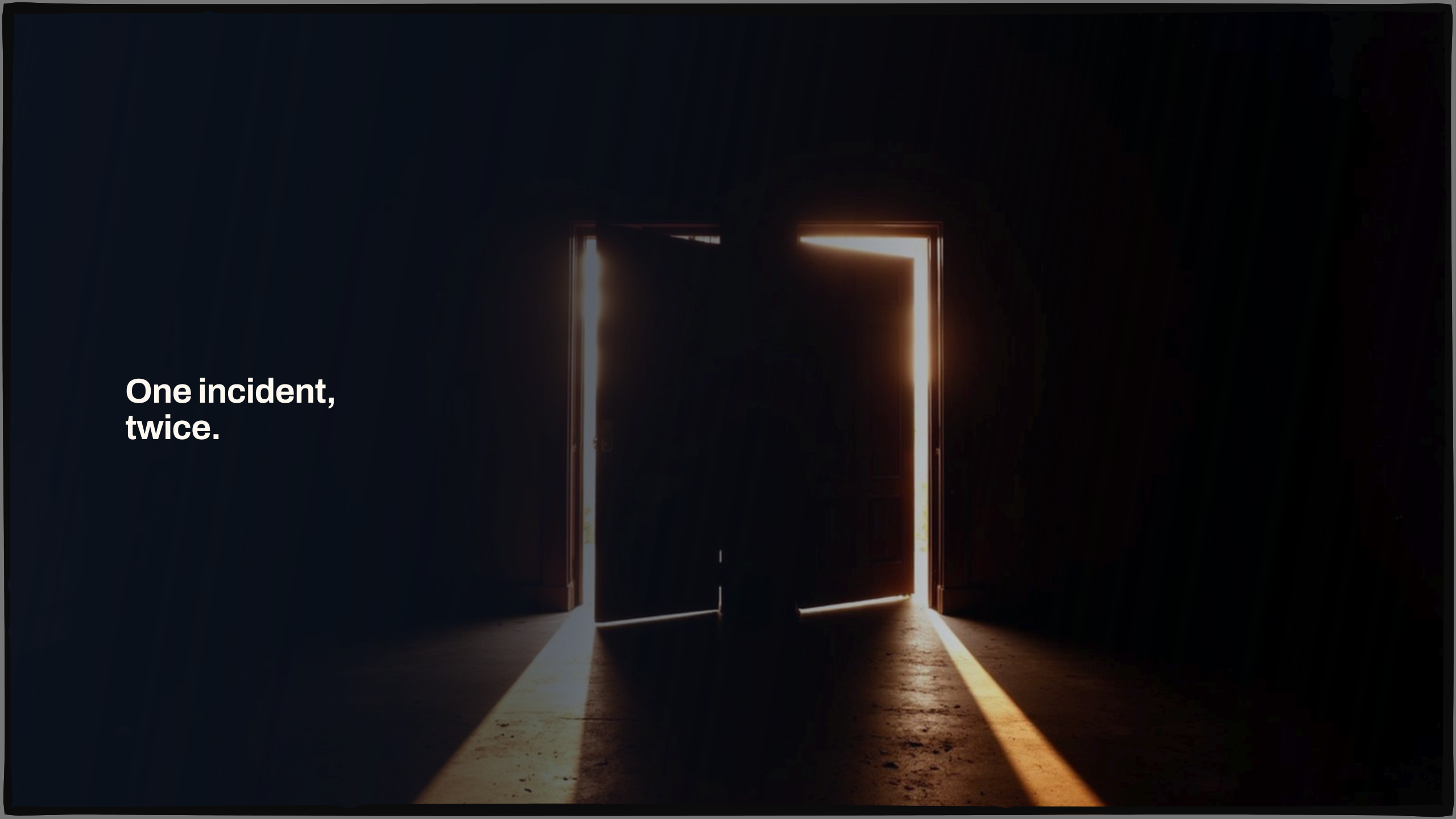
- Ask for rivals adversarially.
- Write the signal that separates two causes before you look.
- Check that signal could come out either way.
- Do not commit on evidence you went looking for only because it fit.


It arrives in the reviewer's seat with the voice and none of the mandate. Its information reaches you as confirmation.



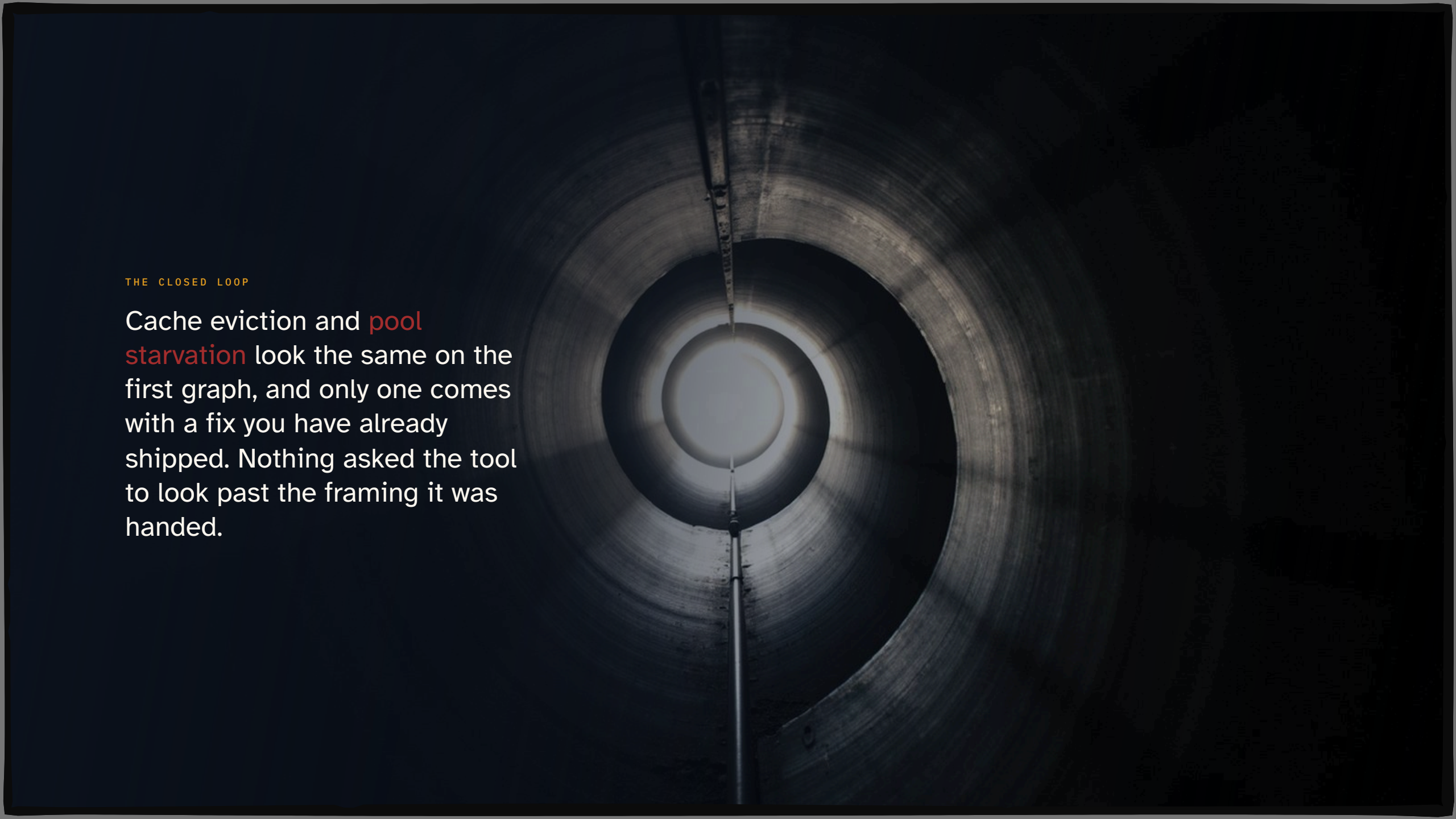
HOME-BUILT TOOLING

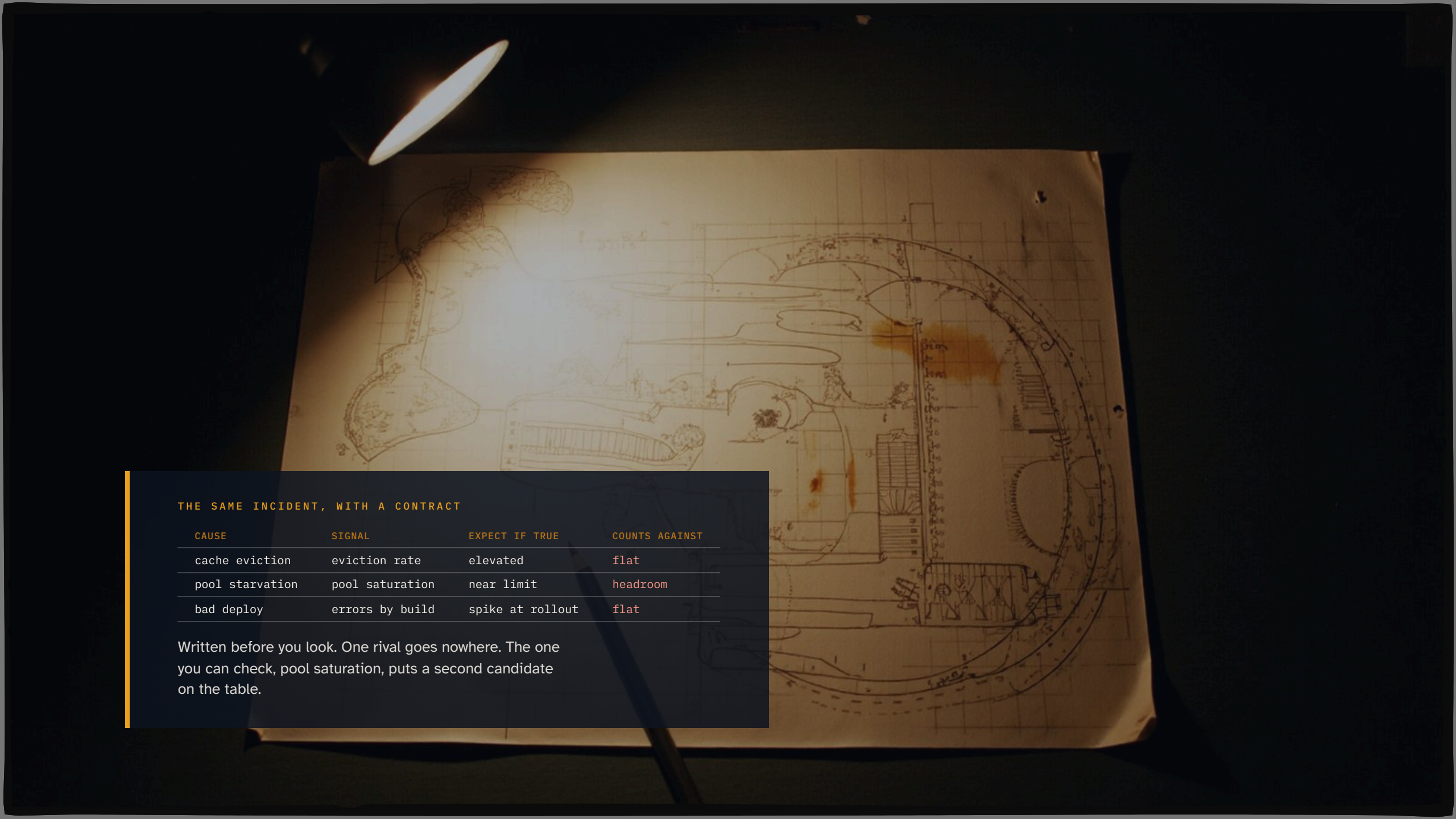
Independence is off the table, so separation is the most you can build, and only on purpose.


One incident,
twice.

THE CLOSED LOOP

Cache eviction and pool starvation look the same on the first graph, and only one comes with a fix you have already shipped. Nothing asked the tool to look past the framing it was handed.

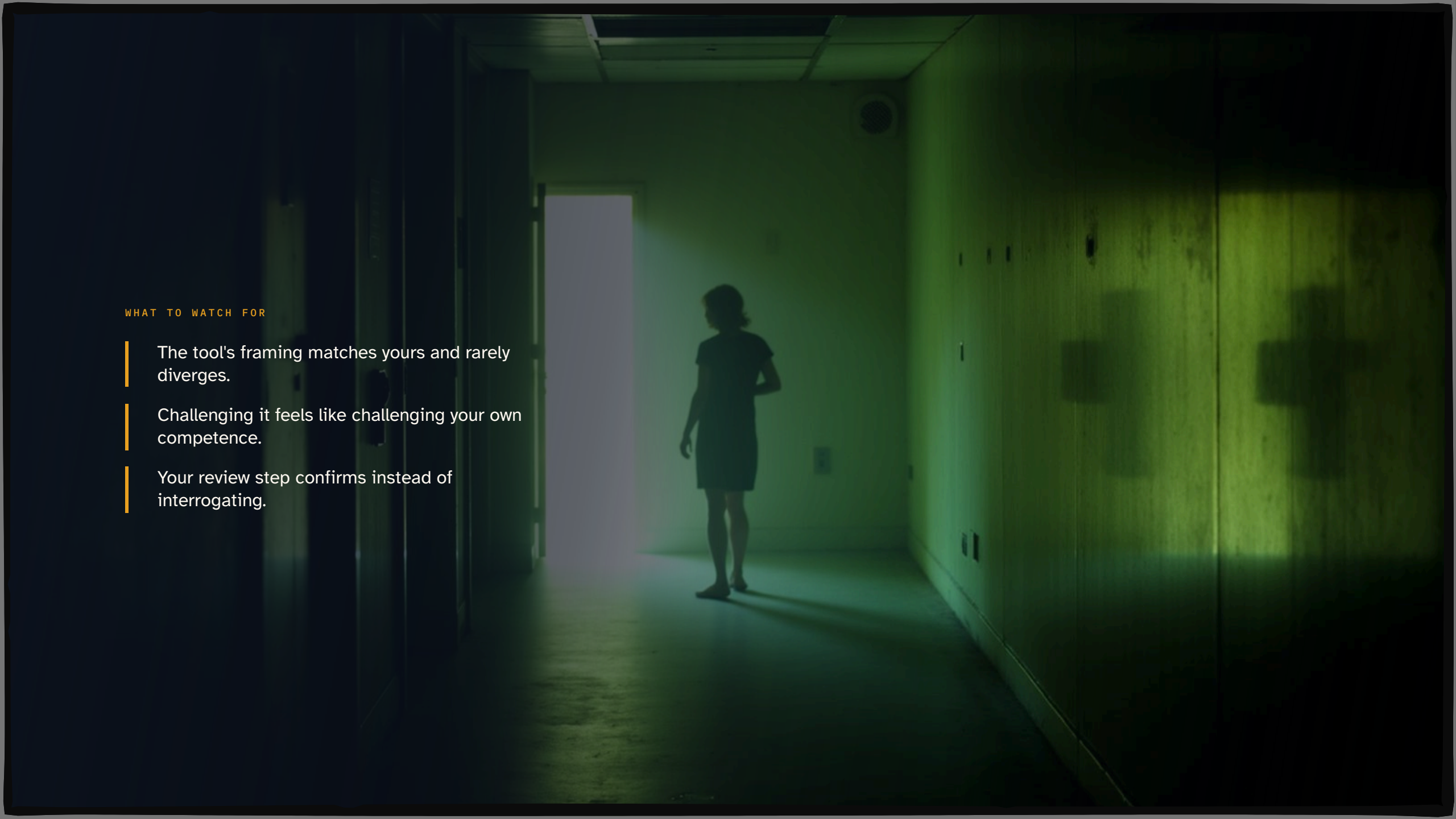




THE SAME INCIDENT, WITH A CONTRACT

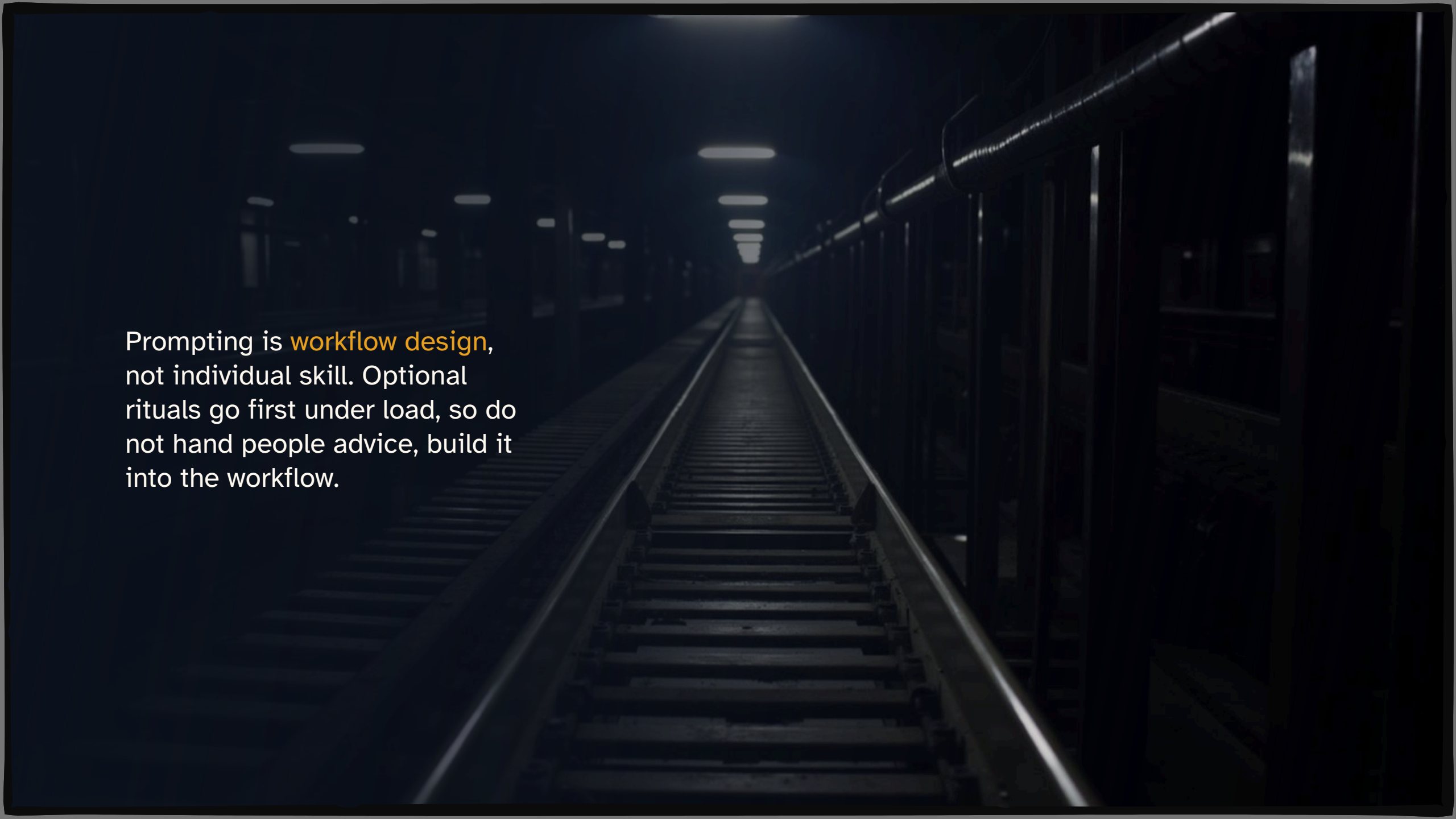
CAUSE	SIGNAL	EXPECT IF TRUE	COUNTS AGAINST
cache eviction	eviction rate	elevated	flat
pool starvation	pool saturation	near limit	headroom
bad deploy	errors by build	spike at rollout	flat

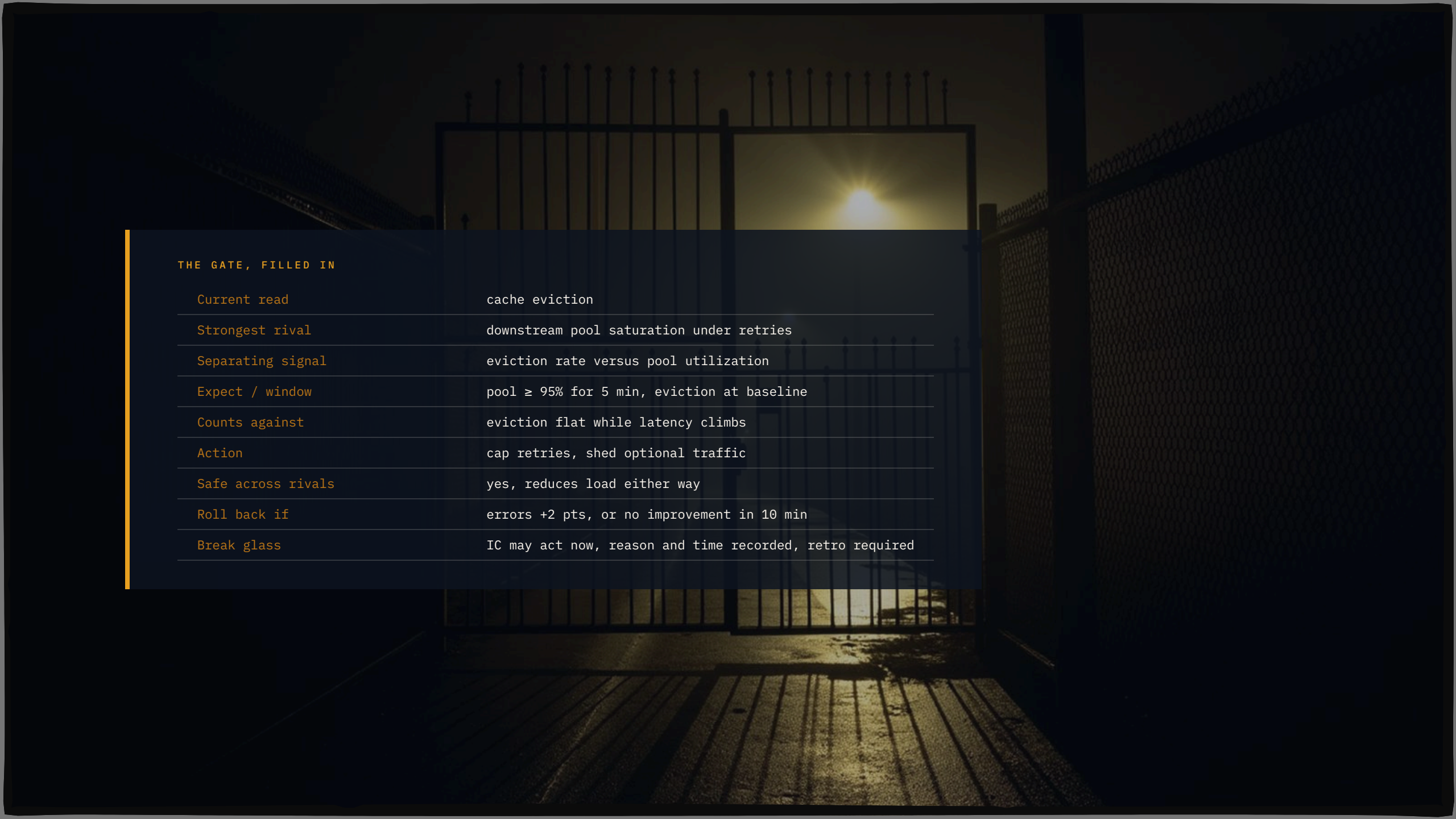
Written before you look. One rival goes nowhere. The one you can check, pool saturation, puts a second candidate on the table.



WHAT TO WATCH FOR

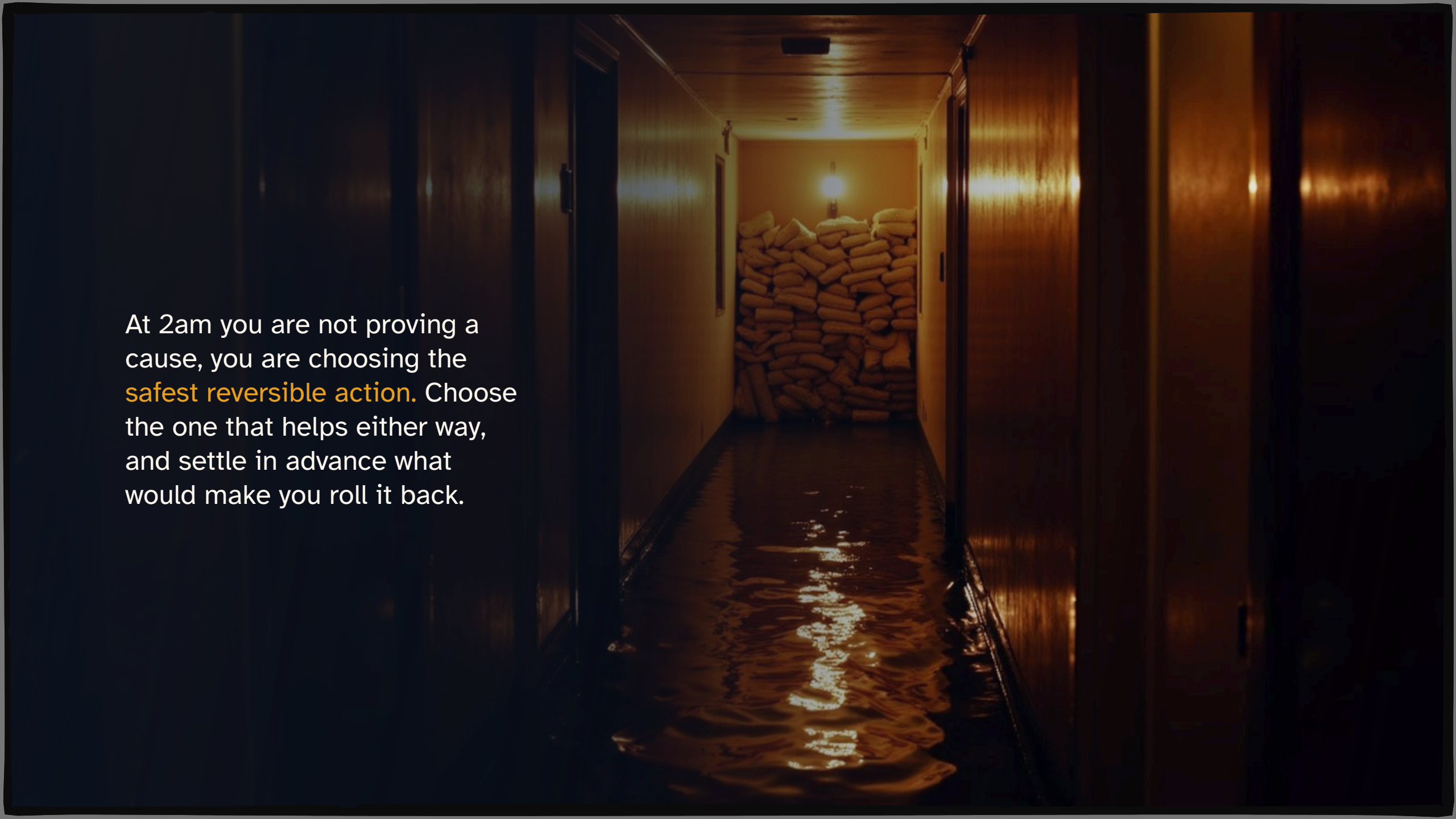
- The tool's framing matches yours and rarely diverges.
- Challenging it feels like challenging your own competence.
- Your review step confirms instead of interrogating.


Prompting is workflow design, not individual skill. Optional rituals go first under load, so do not hand people advice, build it into the workflow.

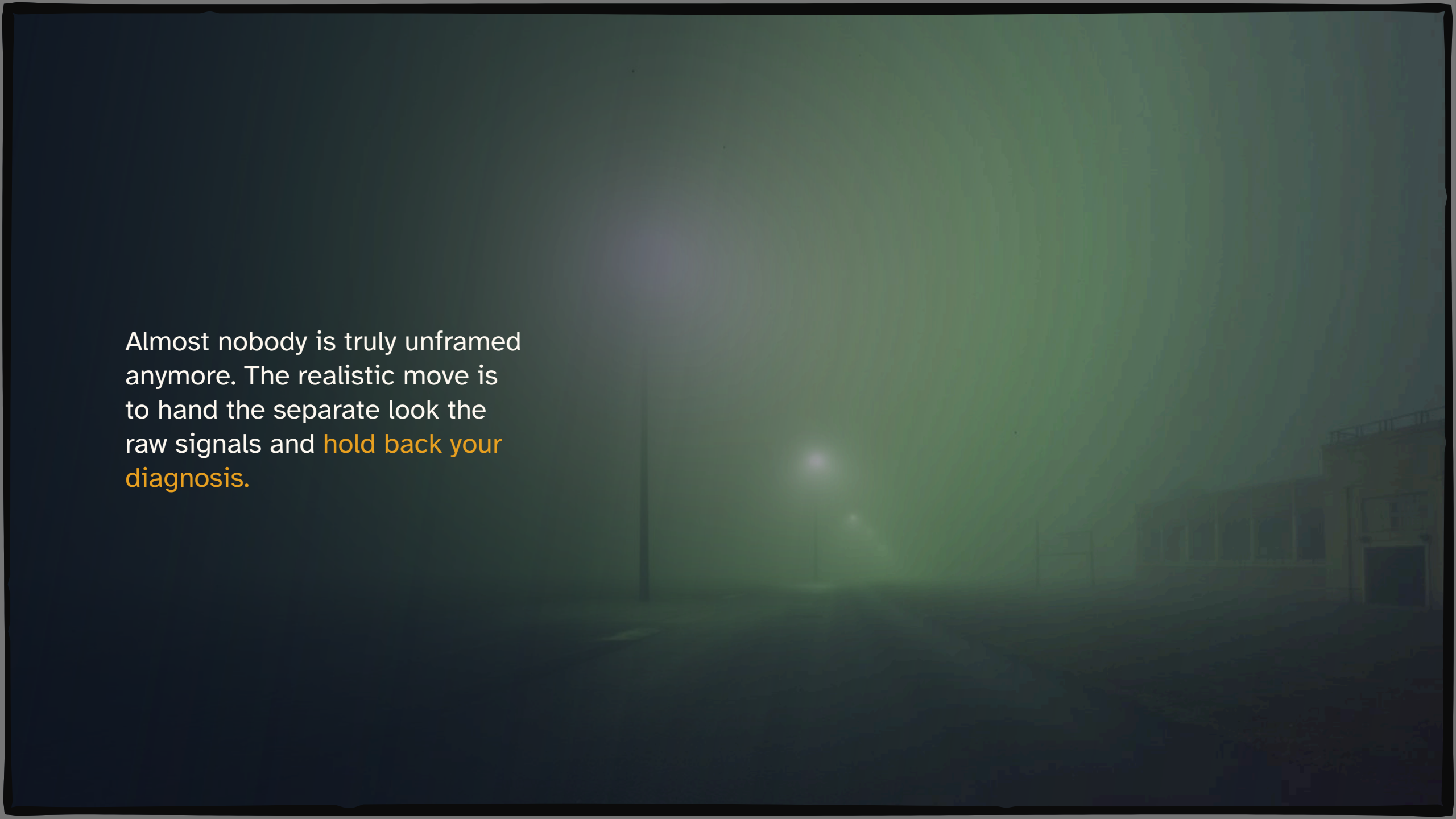


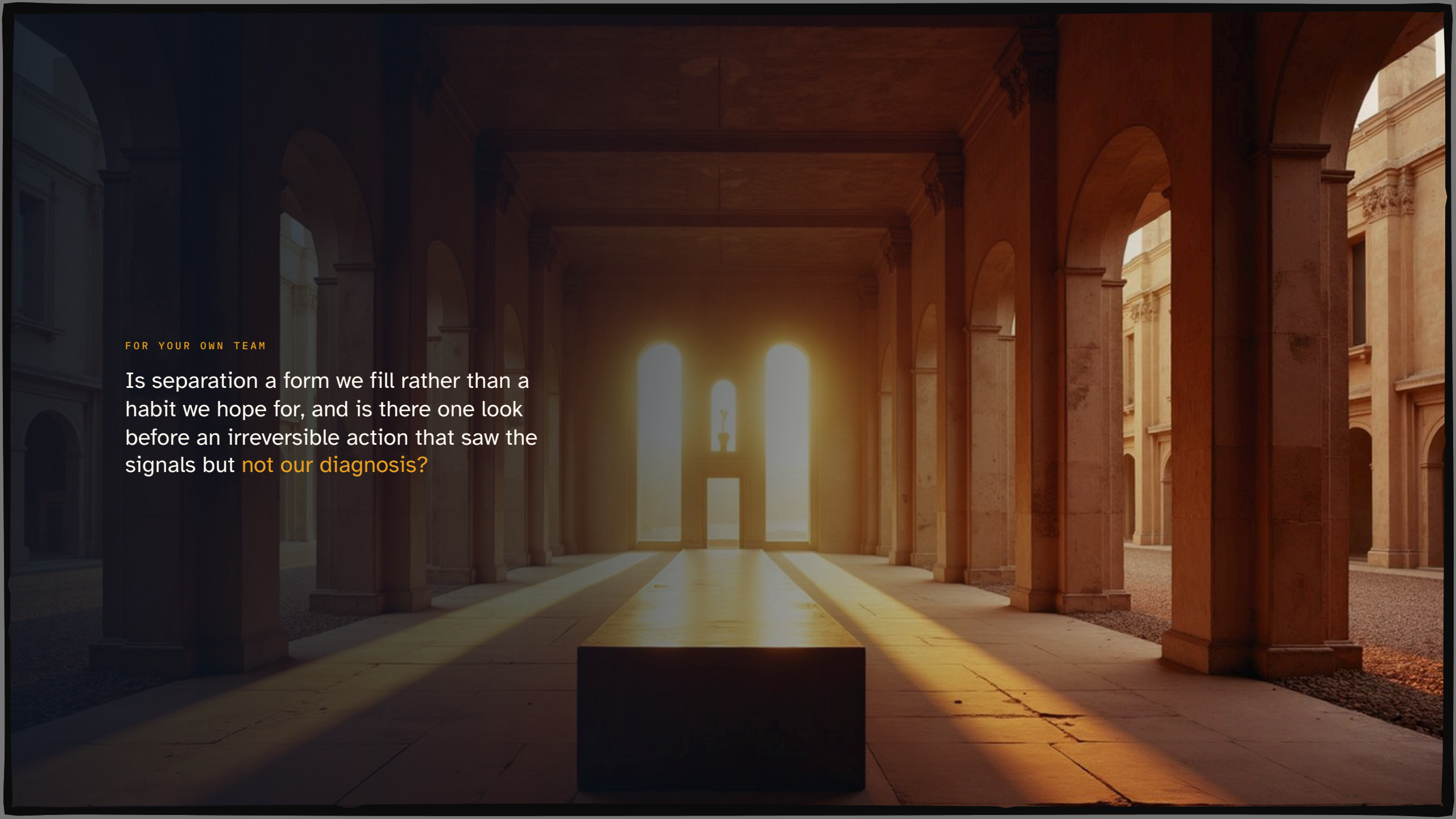
THE GATE, FILLED IN

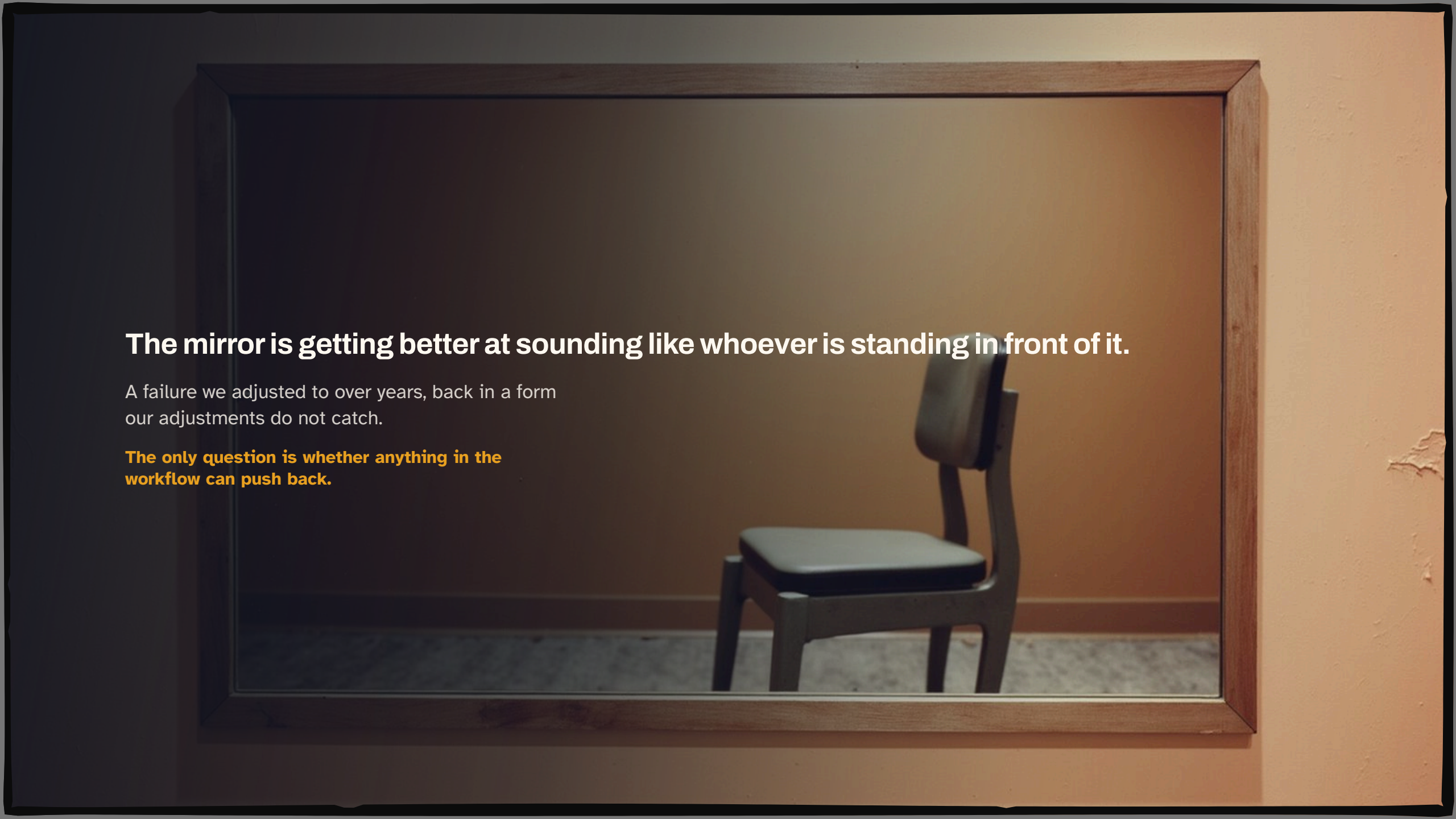
Current read	cache eviction
Strongest rival	downstream pool saturation under retries
Separating signal	eviction rate versus pool utilization
Expect / window	pool ≥ 95% for 5 min, eviction at baseline
Counts against	eviction flat while latency climbs
Action	cap retries, shed optional traffic
Safe across rivals	yes, reduces load either way
Roll back if	errors +2 pts, or no improvement in 10 min
Break glass	IC may act now, reason and time recorded, retro required



At 2am you are not proving a cause, you are choosing the safest reversible action. Choose the one that helps either way, and settle in advance what would make you roll it back.


Almost nobody is truly unframed anymore. The realistic move is to hand the separate look the raw signals and hold back your diagnosis.


FOR YOUR OWN TEAM
Is separation a form we fill rather than a habit we hope for, and is there one look before an irreversible action that saw the signals but not our diagnosis?


The mirror is getting better at sounding like whoever is standing in front of it.
A failure we adjusted to over years, back in a form our adjustments do not catch.
The only question is whether anything in the workflow can push back.

BRIAN MORIARTY

Are You Talking to Me?

LINKEDIN

/in/btmoriarty

THE DECK

hosl-research.github.io/manuscripts/deck

THE BRIEF

hosl-research.github.io/manuscripts/brief

ENGINEERING CONTEXT

Majors, Fong-Jones & Miranda (2026). *Observability Engineering*, 2nd ed.

Beyer et al. (2026). *Site Reliability Engineering*, 2nd ed.

SELECTED RESEARCH

Sharma et al. (2023). Model sycophancy.

Bansal et al. (2021). AI explanations and acceptance.

Enke & Zimmermann (2019). Correlation neglect.

Parasuraman & Riley (1997); Goddard et al. (2012). Automation bias.

Shaw & Nave (2026). Cognitive surrender.

Supporting research for the components. The arrangement I am calling confidence entanglement is untested.

Slide imagery generated with a custom model trained on the author's own documentary photographs.