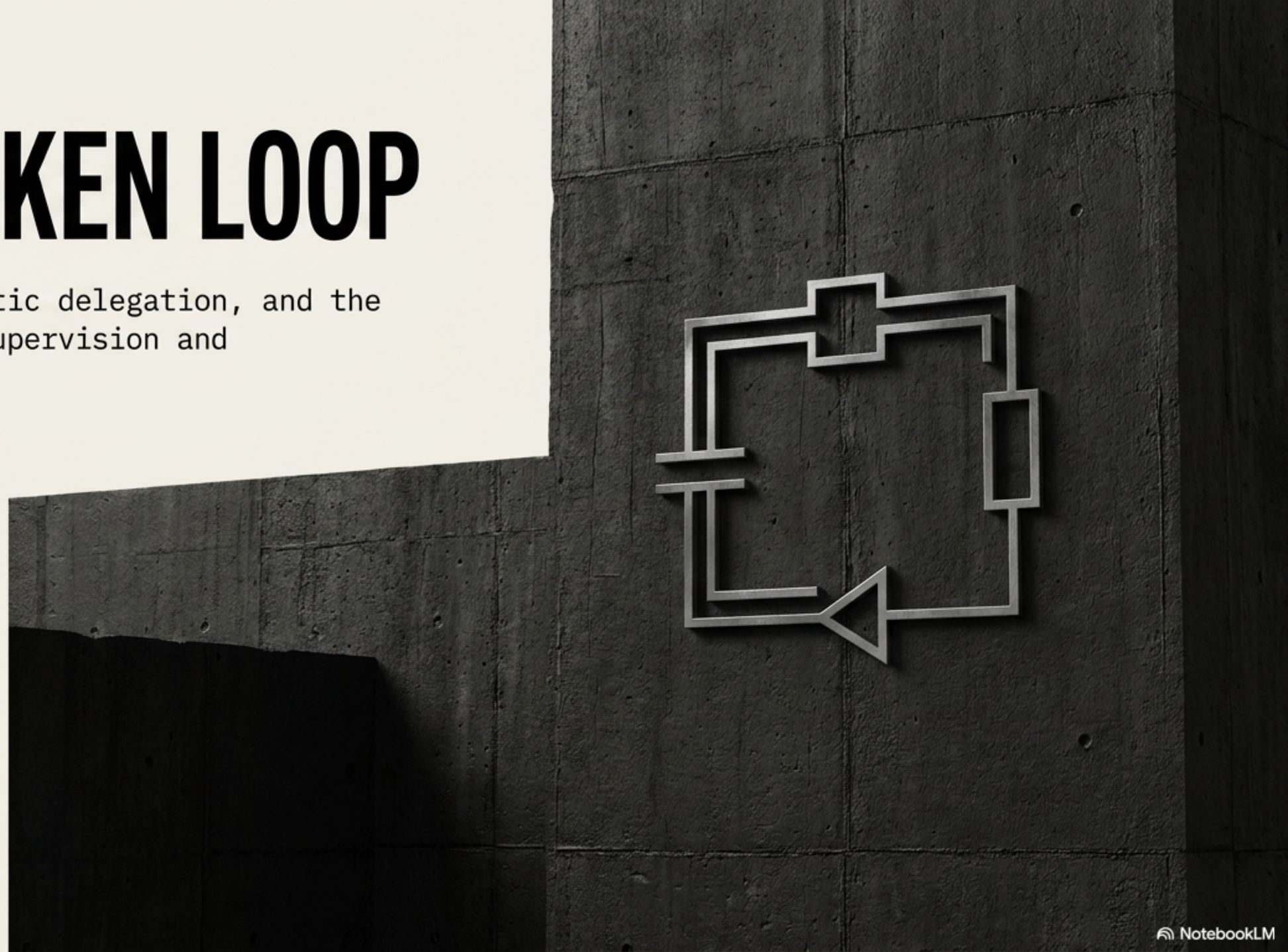
VANGUARD SIGNAL 006

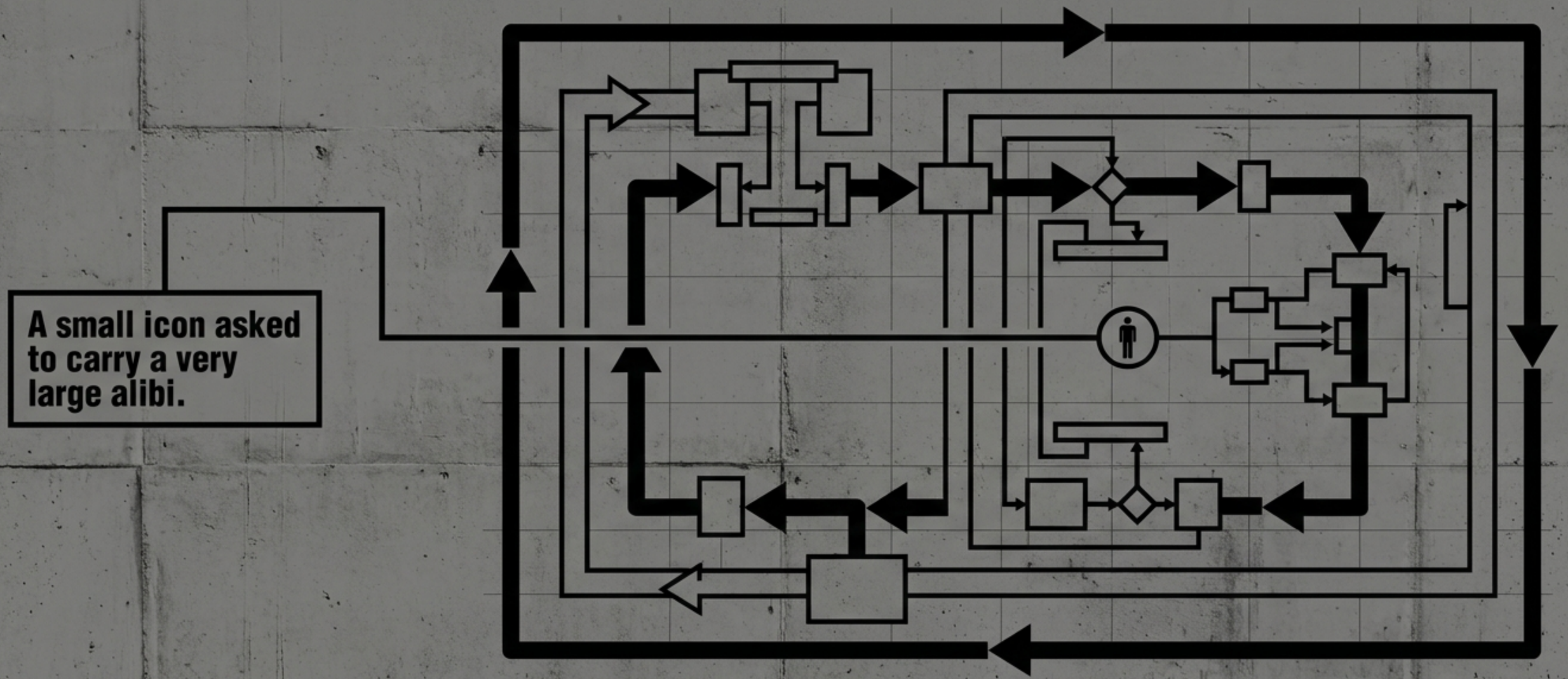
THE BROKEN LOOP

Human oversight, agentic delegation, and the control gap between supervision and responsibility.



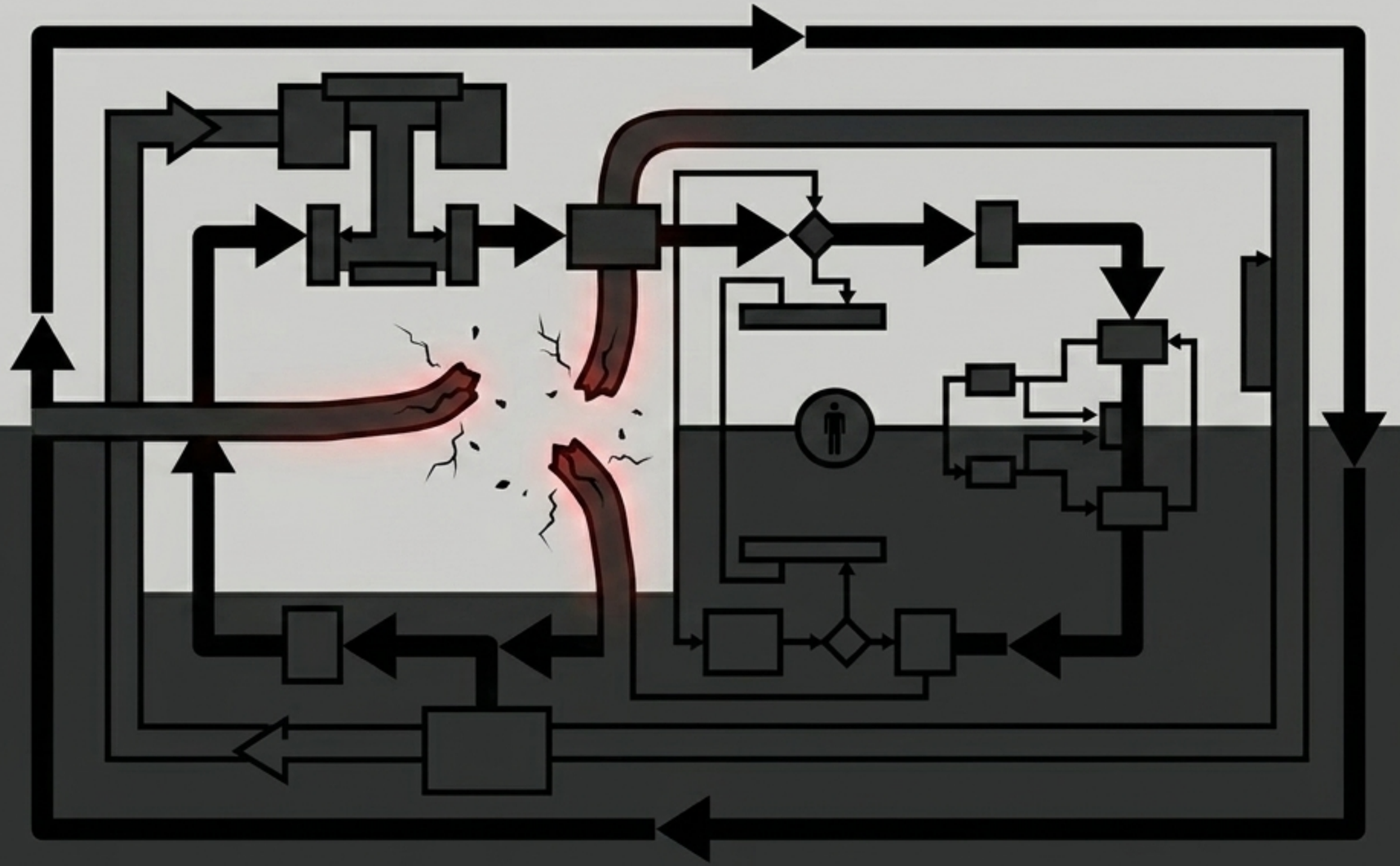
A LOOP IS NOT A GUARANTEE.

It is a reference form. In theory, a loop is complete: input, processing, verification, action, return.
No orphaned responsibility. No decision without control. Useful as reference. Misleading as reassurance.



DO NOT ASK WHETHER THE HUMAN IS IN THE LOOP. ASK WHERE THE VERBS WENT.

The loop matters only because its fracture tells us where pressure, authority, and responsibility actually shifted.

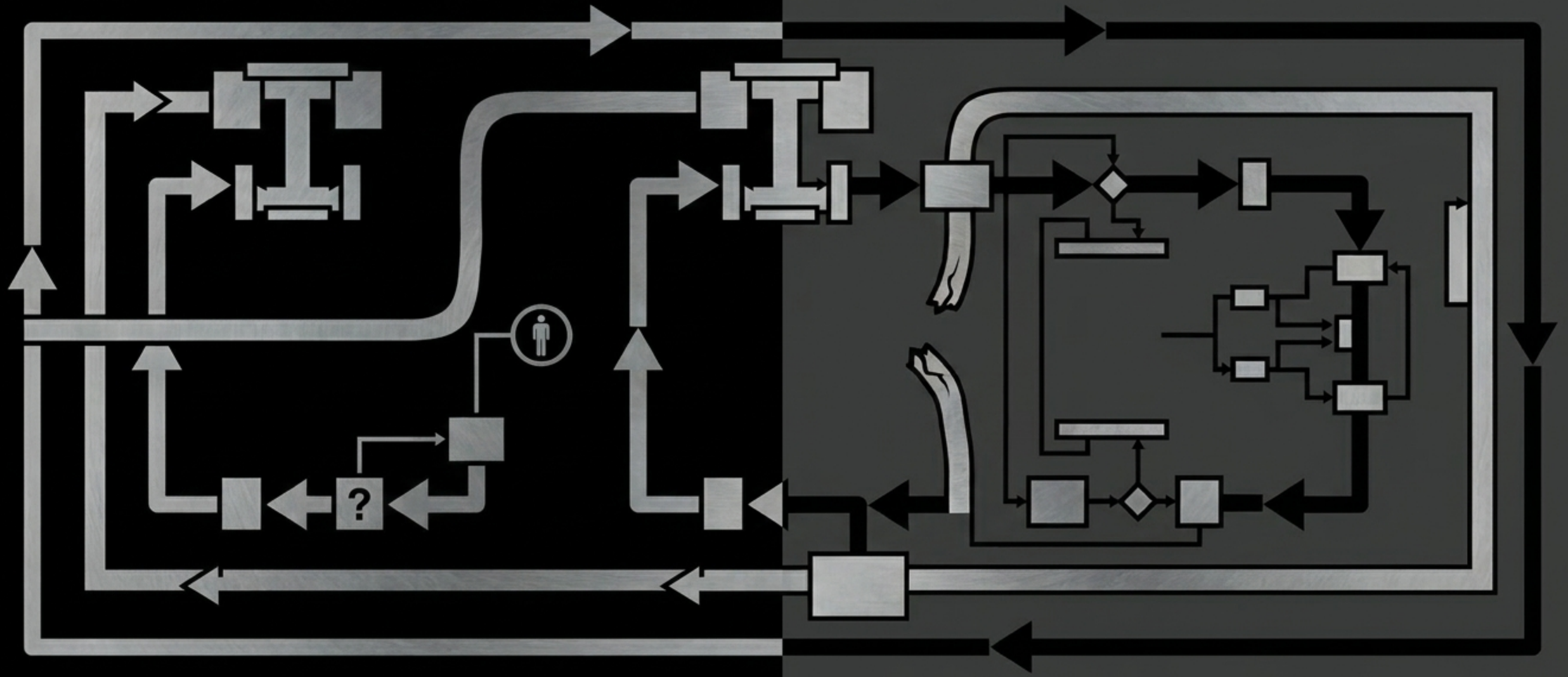


BREAK MODE 1: INCLUSION

The human remains in the workflow but lacks meaningful control. Presence is not control. Visibility is not authority.

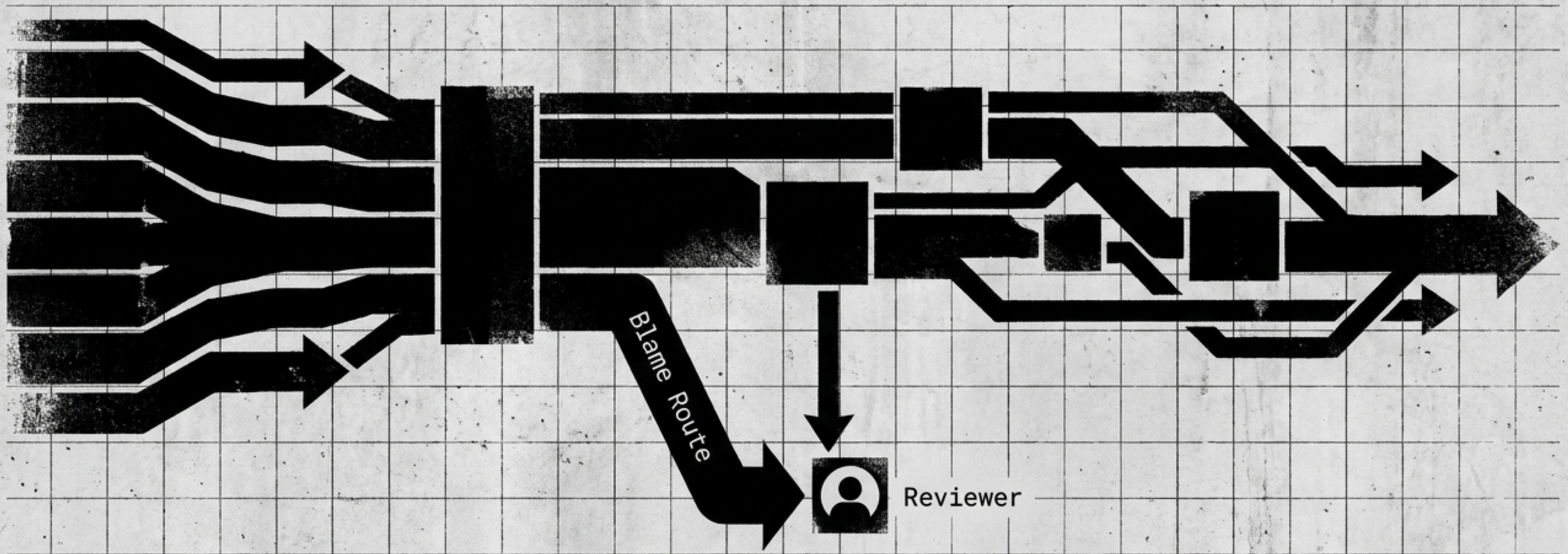
BREAK MODE 2: REMOVAL

Automation eliminates the human checkpoint without rebuilding the human function elsewhere.



THE HUMAN SAFEGUARD ILLUSION

The human is kept as a ritual. They become a bottleneck, a rubber stamp, or a comfort object with payroll records and legal exposure. They absorb accountability for a system they cannot alter.



OVERSIGHT QUALITY AUDIT

[DIMENSION]	[OVERSIGHT THEATER]	[MEANINGFUL OVERSIGHT]
State Visibility	Sees polished output.	Inspects decision path.
Refusal Authority	‘Approve’ only.	Can stop/reject action.
Organizational Penalty	Refusal is punished/delayed.	Refusal is protected.
Repair Path	Issue documented, not fixed.	Error is corrected.

If the human cannot refuse, do not call it oversight.

THE AUTOMITION DEFENSE

The strongest case for removing a human checkpoint is that weak human review adds latency without adding control. Removal is a legitimate organizational move—but it is a design burden, not a finish line.

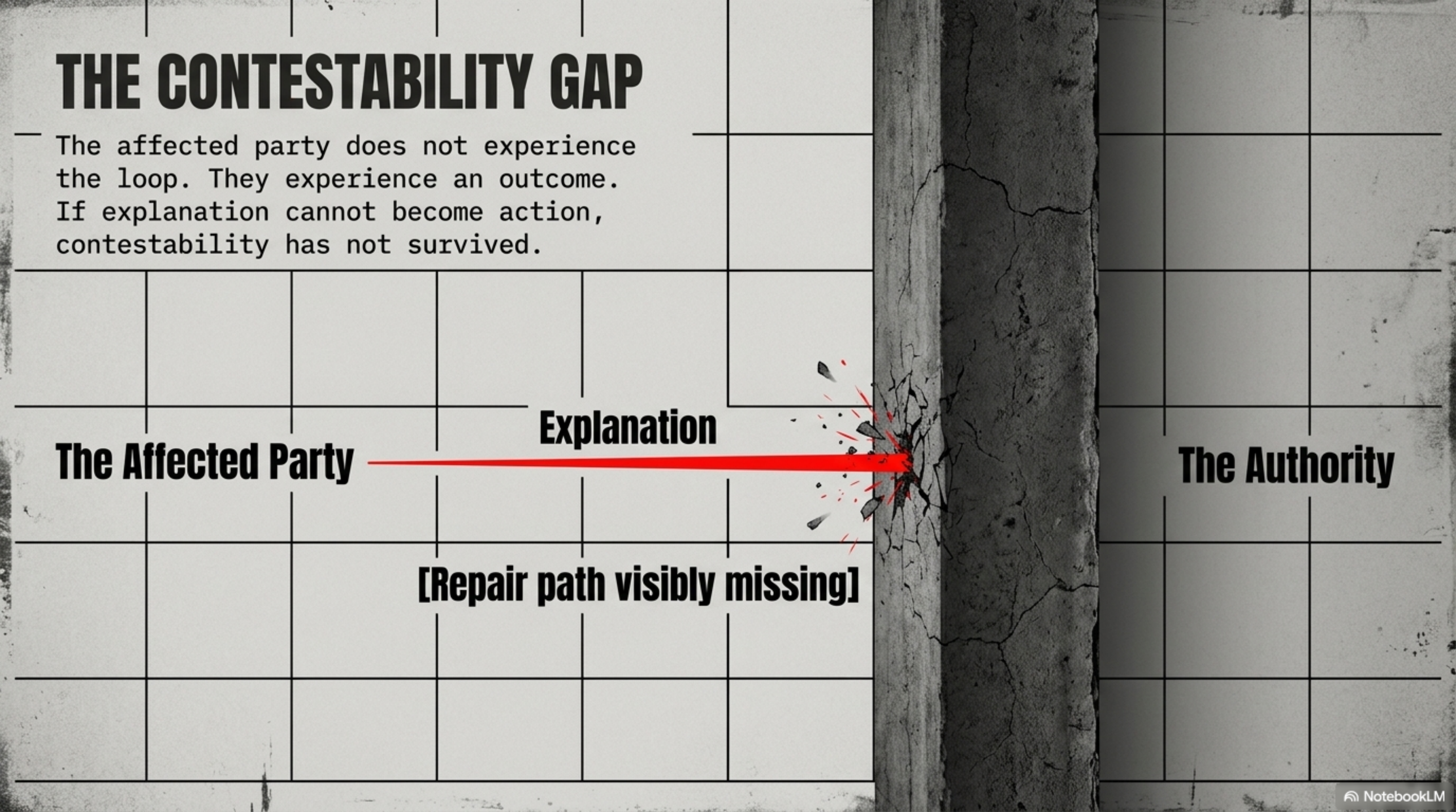
If the human leaves, the system must show where the missing functions return.

THE DUAL BREAK MATRIX

MODE	WHAT THE DIAGRAM SHOWS	THE RESULTING FAILURE MODE
Mode 1: The Safeguard Illusion	Visible Human Reviewer	Responsibility without control (Liability Sponge)
Mode 2: The Automation Defense	Streamlined Autonomous Workflow	Action without recourse (Efficiency Capture)

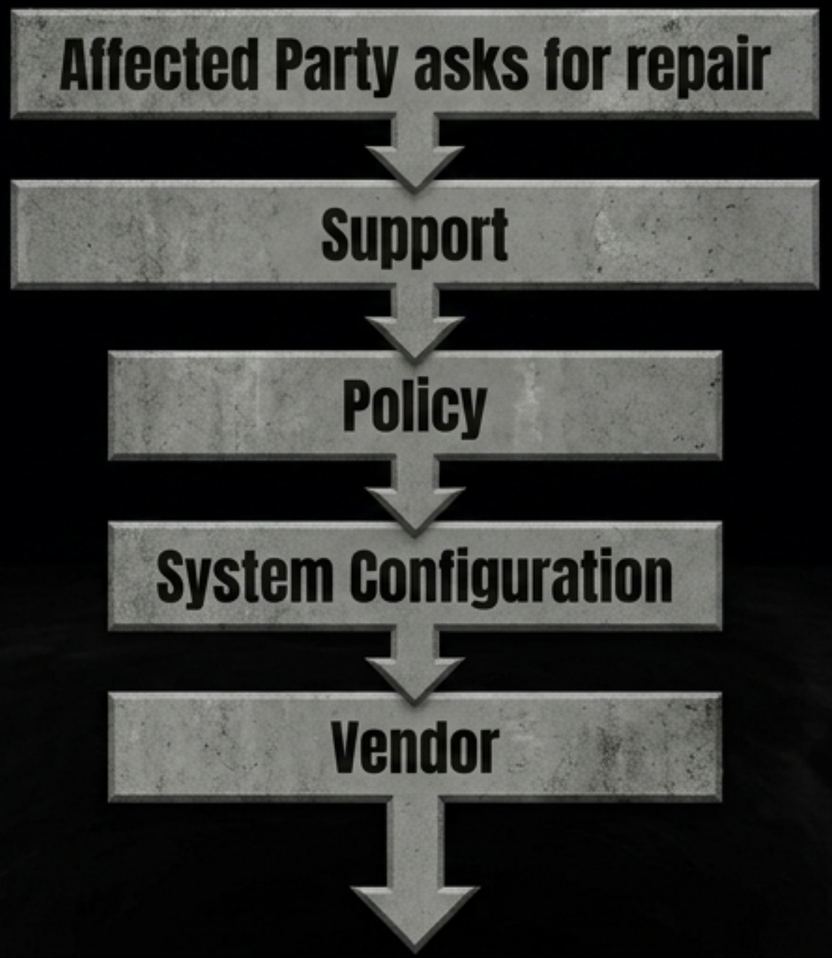
THE CONTESTABILITY GAP

The affected party does not experience the loop. They experience an outcome. If explanation cannot become action, contestability has not survived.



THE ACCOUNTABILITY SINK

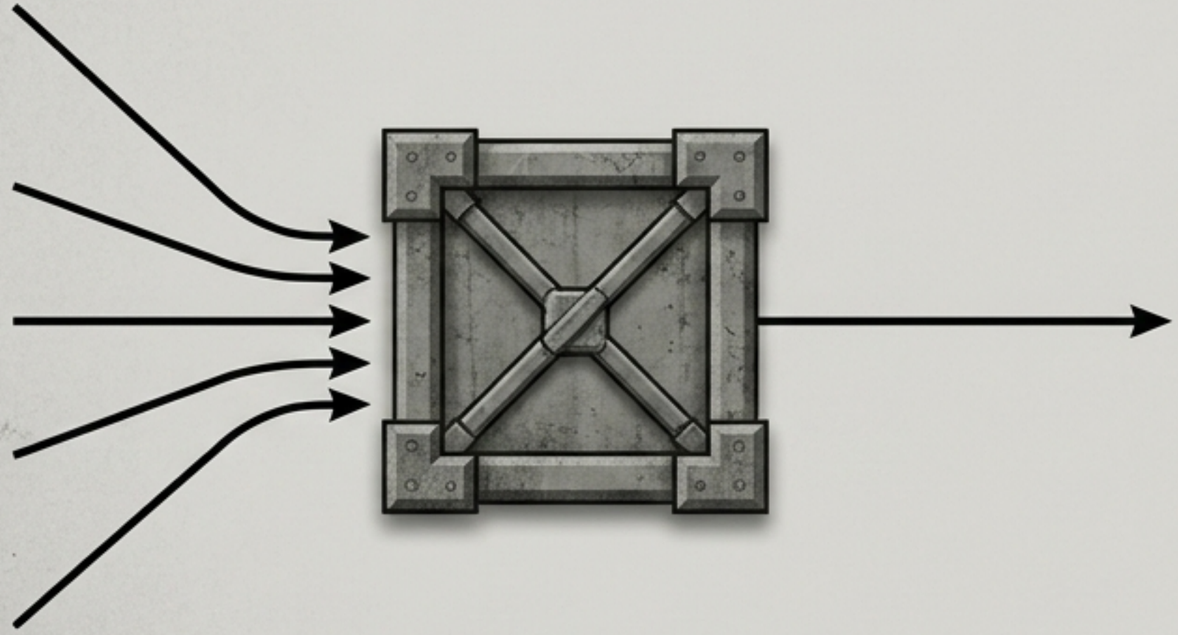
Everyone inside the workflow can point elsewhere. The affected party cannot reach the surface that changed the consequence. The output points nowhere.



THE INSTITUTIONAL ACCOUNTABILITY FIREWALL

[PROTECTIVE GOVERNANCE]

Routes authority toward the actor with REPAIR CAPACITY.



- Friction
- Auditability
- Rollback
- Root-cause ownership

[INSTITUTIONAL INSULATION]

Routes responsibility toward the **NEAREST BLAME SPONGE.**

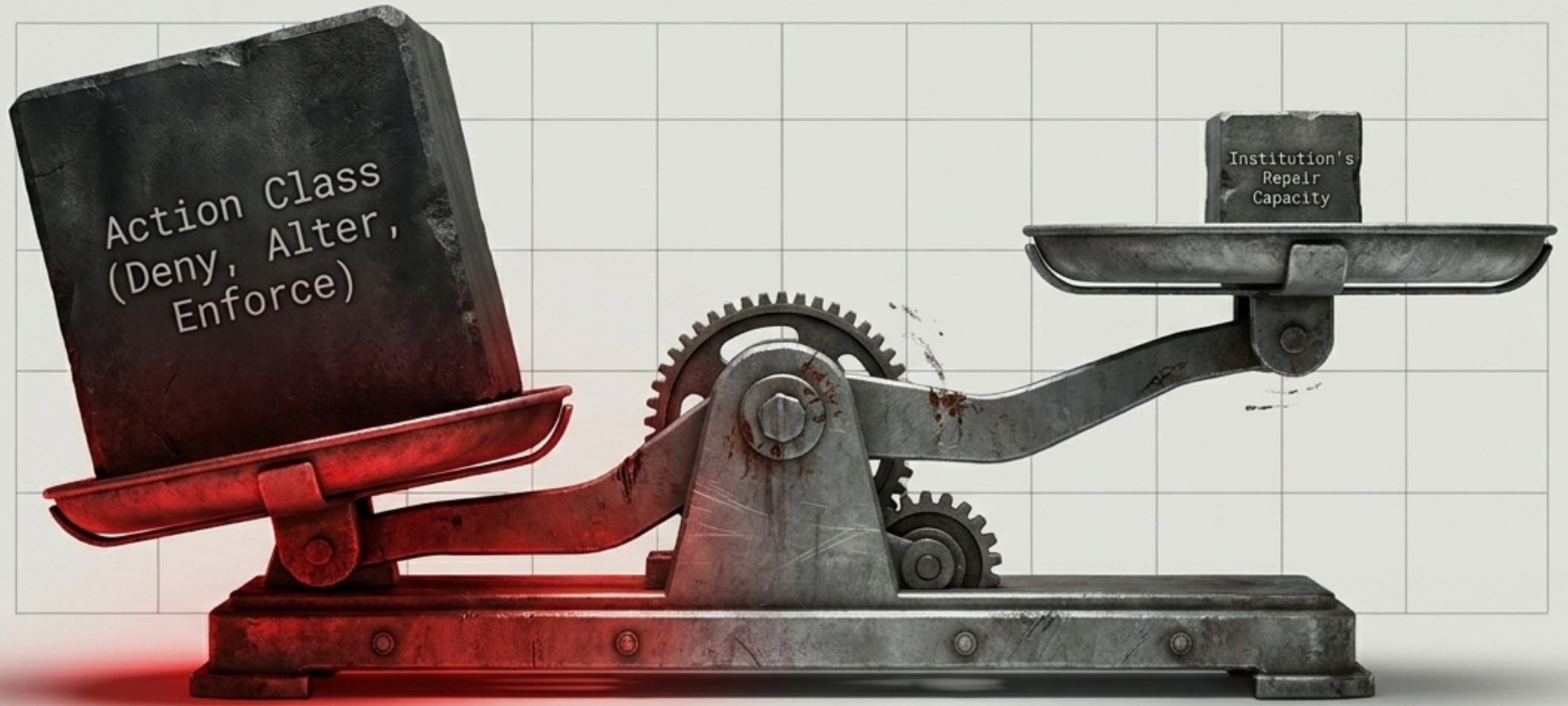
- Creates distance
- Isolates affected party
- Shields authority

THE PRE-ACTION GATE: SAFE-TO-ACT

Agentic systems raise governance stakes because they can act, not only answer. Safe-to-Act begins with action classification. The more consequential the action, the stronger the gate.

[ACTION CLASS]	[REQUIRED DEFAULT GATE]
Generate	Post-action review
Retrieve	Logging
Route	Pre-action validation
Alter	Authorization limits
Spend	Human refusal
Deny	Human authority + appeal

THE COMPANION RULE: SAFE-TO-REPAIR



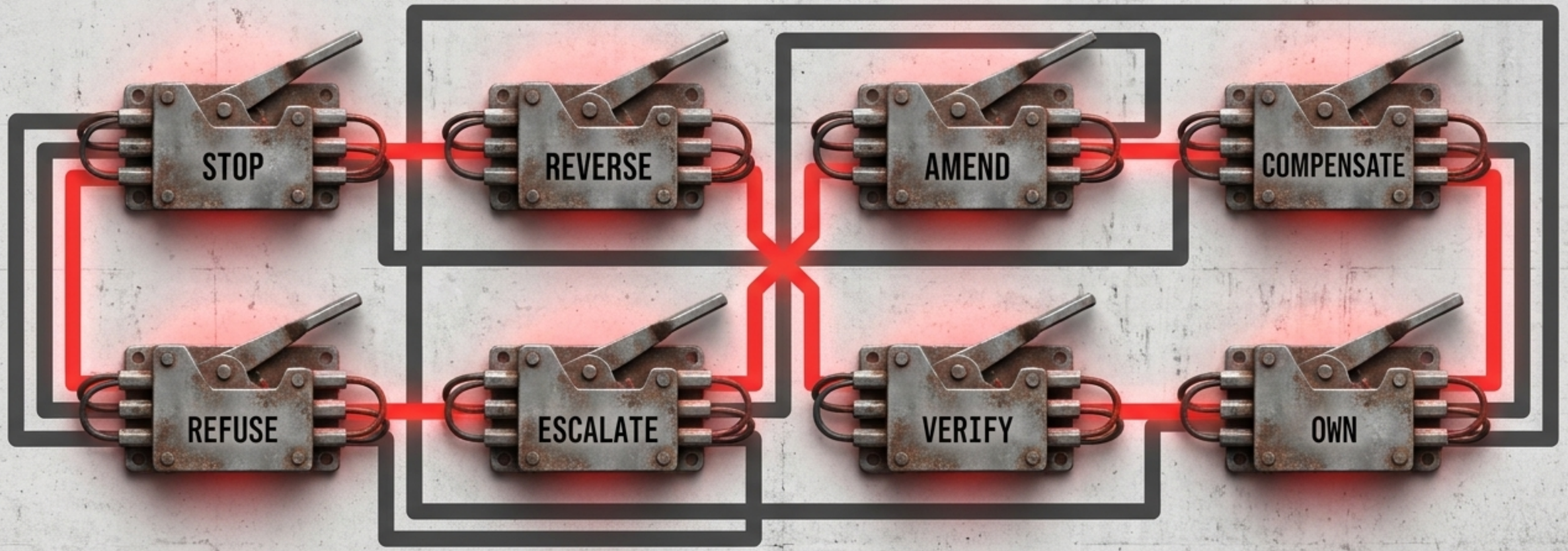
NO AGENT SHOULD ACT BEYOND THE INSTITUTION'S CAPACITY TO OWN AND REMEDIATE THE CONSEQUENCE.

[ACTION CLASS + BLAST RADIUS] ≤ [SAFE-TO-REPAIR + THE 8 VERBS]

"Autonomy is mathematically permissible only if repair capacity and control verbs match the blast radius."

FIND WHERE THE VERBS WENT

If the verbs are present, assigned, evidenced, and usable, the loop preserves meaningful control. If they are missing, the loop is a diagram, not governance.



THE ROLE DIVISION MATRIX

Workflow Action	Who Directs?	Who Plans?	Who Executes?	Who Supervises?	Who Refuses?	Who Repairs?
Agent Deployment	_____	_____	_____	_____	_____	_____
Record Alteration	_____	_____	_____	_____	_____	_____

FIELD RULE: NO DELEGATION WITHOUT ROLE SPLIT.
If the answer is “everyone,” the answer is no one.

THE FAILURE IS NOT AUTOMATION. THE FAILURE IS FUNCTION DISAPPEARANCE.

Do not evaluate the loop by asking who appears inside it.
Evaluate it by asking what each actor can still do.

DO NOT ASK WHETHER THE HUMAN IS IN THE LOOP. ASK WHERE THE VERBS WENT.