No room for slop

Kenneth Pernyér, Reflective Labs
Stockholm, Sweden
kenneth.pernyer@reflective.se · reflective.se · ORCID 0009-0006-4543-1156

My previous report formulated a model of organizational coordination and deliberately stayed above the level of individual people — the target was an initiative's trajectory, never a person's behavior, and the reason given was that behavior is the wrong unit of analysis, not that it doesn't matter. This paper takes up what that one declined to touch. Prediction error, the signal my previous report's whole learning argument depends on, is generated by people — as individuals, whose judgment either engages or quietly disengages when a fluent answer arrives; and in groups, which reliably fail to say what they collectively already know. I formalize both failure classes precisely enough to detect without diagnosing a person: a proposal is slop when its fluency exceeds any trace of the judgment that should have checked it, and a group runs a hidden profile when a fact one member holds is never proposed at all — a failure my substrate's starvation freedom axiom cannot see, because it guarantees a Suggestor executes, not that it discloses. Grounded in the group-decision literature (Stasser and Titus's hidden-profile paradigm, Janis on groupthink, Woolley and colleagues' collective-intelligence factor, Einarsen and colleagues' model of destructive leadership) and in a 2026 finding that extends dual-process theory with a third system — external cognition, and the "cognitive surrender" that follows when it's trusted too readily — I argue these failures are not facts of organizational life to be tolerated. They are failures a sufficiently well-built ever-learning organization has no room for, in the same sense my first report's invariants make certain system states unrepresentable rather than merely discouraged. What that requires, and what it explicitly does not license, are both stated plainly.

1. Introduction

My first report in this series argued that an organization's commitments are bounded by the quality of its shared understanding, and that the understanding only updates when something in it turns out to be wrong — when expectation and outcome come apart. It formulated a model aimed at exactly that gap, at the level of an initiative's trajectory, and it was careful to say why the unit of analysis had to stay there: a model that scores a person instead of a coordination state is a categorically different, more dangerous thing to build (Pernyér, 2025a).

That carefulness left a real question unanswered. Prediction error doesn't generate itself. People generate it — or fail to, which is the harder case. An individual whose judgment has quietly stopped engaging produces fluent output with nothing behind it. A group that has correctly gathered a room full of relevant knowledge can still fail, systematically, to say the part of it that would have changed the answer. Both are coordination failures in the sense my first report cared about — they degrade the shared understanding commitments are bounded by — and neither is visible from the trajectory data alone, because both can look, from the outside, exactly like a decision that was properly made.

This paper takes up what the first one declined to. Not by scoring people — the discipline that report established holds here without exception — but by naming, precisely enough to detect, two failure classes: what an individual does when a fluent answer arrives before judgment has engaged, and what a group does when it has the right people in the room and still doesn't hear what they know. §2 formalizes the individual case. §3 formalizes the group case. §4 connects both back to the substrate — where the promotion gate already covers this and where it structurally cannot — and states the paper's actual claim: not that these failures should be watched for, but that an organization built the way this series has been arguing for has no room for them to persist. §5 states what that claim does not license.

2. The individual case: cognitive surrender

Daniel Kahneman's account of judgment distinguishes fast, intuitive processing from slow, deliberate reasoning — System 1 and System 2 (Kahneman, 2011). A 2026 study extends that account with a third system: external cognition, artificial intelligence operating alongside the other two rather than replacing either (Shaw and Nave, 2026). Across three preregistered experiments — nearly 1,400 participants, close to ten thousand trials — the study's central finding

is that people consult external cognition on a majority of judgment tasks, and that doing so measurably changes the accuracy of their conclusions in both directions: sharply better when the external system is right, sharply worse when it's wrong, and worse in proportion to how readily the person accepted its output without independently checking it. The authors name this **cognitive surrender**: adopting an external system's conclusion with minimal scrutiny, in a way that overrides both the intuition and the deliberation that would otherwise have caught an error.

> **Definition (Slop).** A proposal is **slop** if its fluency — how complete and confident it reads — exceeds any trace of the judgment that should have checked it before it was offered as a conclusion. Slop is not a claim about whether the proposal is wrong. A fluent, unchecked proposal can happen to be right. It is a claim about whether anyone's judgment actually engaged before it was asserted.

This is where the connection to my first report's substrate becomes precise rather than analogical. That report's promotion gate checks a proposal's authority, its schema, and its confidence before admitting it as a fact (Pernyér, 2025a). None of the three checks that gate performs can currently tell the difference between a human's genuine approval and a human's cognitive surrender — both produce an identical record: a proposal, a named approver, a promoted fact. The gate was built to keep an unchecked machine proposal from becoming a commitment. It has no equivalent defense against an unchecked **human** proposal, where the human's approval is real in the sense that a person's name is attached to it, and unreal in the sense that no System 2 judgment actually ran.

> **Open problem (Surrendered approval).** Give the promotion gate a way to distinguish a proposal a human approved after genuine deliberation from one a human approved after cognitive surrender, given that both currently produce the same recorded artifact: a proposal, an approving identity, a promoted fact.

I don't solve this here. What follows from stating it precisely is a design constraint rather than a mechanism: anything that claims to detect cognitive surrender has to look for the **absence of engagement** — a missing counter-question, an approval with no elapsed time between the proposal arriving and the approval being recorded, a pattern of accepting a system's first answer regardless of what that answer says — never for a trait of the person doing the approving. The failure being named is a gap in a process. It is not a character judgment, and a detector that reads like one has already violated the safe-output discipline my first report established.

I'll call the alternative — a person's judgment staying structurally engaged even where an external system produces a fluent, complete-looking answer — **System 2b**: not a rejection of external cognition, a specific discipline for using it, in which the system is required to sharpen the question before it's trusted to have answered it. An organization made of surrendered approvals accumulates commitments nobody actually checked. An organization made of System 2b accumulates commitments that were genuinely tested, whether or not an external system was involved in reaching them. The difference doesn't show up in the commitment record. It shows up, if anywhere, in whether a commitment survives contact with reality — which is exactly the prediction error my first report's whole learning argument runs on. Slop doesn't just produce a bad individual decision. It quietly degrades the organization's capacity to generate the one signal it can actually learn from, because a surrendered approval isn't a judgment that turned out to be wrong. It's not a judgment at all, and there's nothing in it to learn from when it fails.

3. The group case: what the room already knew

A single person's cognitive surrender is a failure of one judgment. A group's failure is different in kind, and older in the literature: a group can be made of people whose judgment engaged individually and still reach a worse conclusion than any one of them would have reached alone, because of how the group combines what its members know.

3.1. The hidden profile

The clearest version of this was demonstrated four decades ago and has held up since. Give a group of people a decision with a correct answer that only becomes visible if the group pools everyone's information — some facts known to everyone, others known to only one person — and groups reliably fail to surface the facts only one person holds. They spend the discussion on what everyone already agreed on, converge on the option the shared information favors, and never reach the option the **unshared** information would have favored, even though every piece of it was sitting in the room the whole time (Stasser and Titus, 1985).

> **Definition (Hidden profile).** A group has a hidden profile on a decision if some member holds evidence that would change the group's conclusion, and that evidence is never proposed into the group's shared consideration — not rejected, not contested, simply never said.

This connects to my first report's substrate at a specific, sharp point, and it's worth being precise about exactly where. Starvation freedom — one of the nine invariants that report establishes — guarantees that every eligible Suggestor eventually executes (Pernyér, 2025a). Applied to a human participant in a decision process, that guarantee is real: being asked, being given the floor, being on the list of people consulted. What starvation freedom cannot guarantee, and was never built to, is that a Suggestor who executes actually **discloses** what it holds. A person can be asked, can speak, and can still not say the one thing that would have changed the outcome — not from malice, not from being silenced, but because the group's discussion pattern never surfaced it as relevant. Execution is a property the substrate can enforce. Disclosure is not, and the hidden-profile literature is precisely the description of how a group of individually well-functioning members produces a collectively worse answer through exactly this gap.

3.2. Groupthink and the cost of visible agreement

A related and equally well-established finding runs the other direction: a sufficiently cohesive group under pressure to agree will actively suppress dissent rather than merely fail to surface it, producing high collective confidence in a conclusion that a less cohesive group would have caught (Janis, 1972). The two failures compound rather than cancel. A hidden profile means the disconfirming fact never entered the room. Groupthink means that even facts that **did** enter the room get discounted once the group has converged, because raising them now reads as friction rather than information. Agreement, under either failure, stops being evidence that the group reasoned well and becomes evidence that the group stopped checking.

3.3. HiPPO, named informally and grounded precisely

The pattern where a room converges on the view of its most senior or highest- status member, independent of the evidence, has an informal name in analytics and product culture — HiPPO, the highest-paid person's opinion — and I use that name because it's the one practitioners will recognize, not because a single paper coined it. What it names is real and well studied: status and authority measurably shift what a group is willing to say and which contributions it credits, independent of the contribution's actual merit, and a group whose contributions correlate with rank rather than with information has degraded in exactly the way a hidden profile does, for a more visible reason. Where HiPPO and hidden profile differ is in what they need as a remedy. Hidden profile needs a structural prompt: ask explicitly what each person holds that others might not. HiPPO needs a structural constraint: a contribution's standing in the record has to be independent of who made it, checked, not assumed.

3.4. Distinguishing a toxic pattern from a toxic person

None of this paper's discipline is more load-bearing than it is here. The organizational-leadership literature has a precise model of what makes a leader's behavior destructive: systematic, repeated conduct that undermines the organization's own goals and its people's wellbeing, falling into recognizable patterns — tyrannical conduct toward subordinates, leadership that has derailed from the organization's actual interests, and a supportive-disloyal pattern that protects a team while working against the wider organization (Einarsen et al., 2007). That taxonomy is exactly the register this paper's safe-output discipline requires: it names **behavior patterns**, each falsifiable against a record of specific incidents, not a diagnosis of a person's character. "This manager exhibits a pattern consistent with tyrannical leadership behavior — six documented incidents of overriding team decisions without stated reasoning in the last quarter, each contested internally and none revised" is a structural claim with cited evidence. "This manager is toxic" is a judgment the model has no business making, for the same reason my first report ruled out "Alice is disengaged." The taxonomy gives the first sentence a vocabulary. It does not license the second.

3.5. What good groups do instead

The failure modes above have a positive complement worth naming, because a detector built only to catch failure has no target to aim an organization toward. Group performance across a wide range of tasks correlates weakly with the individual intelligence of a group's members and strongly with two measurable properties of how the group interacts: the average social sensitivity of its members, and how evenly participation is distributed rather than concentrated in a few voices (Woolley et al., 2010). Neither property is about any individual's talent. Both are about the structure of the interaction — which is, again, exactly the register a coordination-focused model is allowed to operate in.

4. Where the substrate already helps, and where it has to

My second report described a mechanism that addresses part of what this paper has named, built before this paper existed and not built for this purpose: institutionalized disagreement, in which every plan is challenged by assumption-breaking, constraint-checking, causal, economic, and operational skepticism before it may proceed, with a blocker acting as a veto no quorum can overrule (Pernyér,

2025b). That report showed why the veto has to be structural rather than a vote — a majority of skeptics can each individually believe a plan is unsound while a premise-by-premise vote passes it regardless, a result that generalizes past adversarial review to any process that aggregates judgments by counting rather than by testing them. Mandatory, structural, adversarial challenge is precisely the mechanism that makes both groupthink and HiPPO harder to sustain: groupthink because dissent is no longer optional and its absence is no longer read as evidence of soundness; HiPPO because a senior participant's plan faces the same five skepticisms as anyone else's, with no exemption the record would show.

What that mechanism does not yet reach is the hidden profile and cognitive surrender specifically, because both are failures of **disclosure** rather than failures of **challenge**. Adversarial review tests a plan that has already been stated. It has no lever on the fact that never got stated at all, or on the approval that was never really deliberated before it was given. Closing that gap needs a different kind of structural prompt than a skeptic asking what's wrong with a stated plan — it needs something that asks each participant, before convergence, what they know that the group's current direction doesn't yet reflect, and something that can tell a genuinely tested approval from a surrendered one. Neither exists in this substrate today. Naming the gap precisely is this paper's actual contribution; closing it is not attempted here.

5. What this does not license

Four things are worth stating plainly, because a paper that names failure patterns this specific has to be equally specific about where it stops.

This paper does not propose scoring, ranking, or flagging individual people. Every definition above is stated over **proposals**, **approvals**, and **disclosures** — artifacts of a process — never over a person's trait or character. A system that reduces any of this to a person-level score has violated the paper's own discipline, not extended it.

The destructive-leadership taxonomy in §3.4 names patterns strong enough to require a real evidentiary standard before they're asserted about anyone — repeated, documented incidents, not a single flagged interaction — and this paper does not specify what that standard should be. That's a governance question, not a modelling one, and it belongs with whoever is accountable for the consequences of getting it wrong, not with whatever detects the pattern.

Cognitive-surrender detection at the individual level is the most sensitive claim in this paper and the least developed. The open problem in §2 is stated as a problem, not a design. I have no evidence that the pattern it describes is reliably distinguishable from ordinary, well-justified trust in a system that has earned it, and treating a fast approval as evidence of surrender rather than of competence would be exactly the kind of unsafe inference §2 was written to rule out.

Finally, this paper inherits every limitation my first report already stated about the substrate it depends on — the data this would require, the provenance discipline any detector would have to meet, the postponement of actual training. Nothing here changes that status. This is a formulation of what the human layer of an ever-learning organization would have to account for, not a system that accounts for it yet.

6. Conclusion

My first report was about the downstream symptom of a shared understanding degrading — a commitment outliving the conditions that justified it. This one is about two of the upstream mechanisms by which that degradation actually happens: a person's judgment quietly disengaging when a fluent answer arrives, and a group failing to say what it collectively already knows. Both are old problems, with real literatures behind them, and both turn out to connect precisely to invariants this series has already proven — the promotion gate that can't yet tell a genuine approval from a surrendered one, starvation freedom that guarantees a voice gets heard and says nothing about whether it speaks.

An organization built the way this series argues for is not one that merely watches for these failures. It's one where institutionalized disagreement already makes groupthink and HiPPO harder to sustain, where the remaining gap — disclosure, not challenge — is named precisely enough to close, and where closing it never means scoring a person, only naming a pattern in a process and pointing back at the evidence. That is what "no room for slop" is actually claiming: not vigilance, structure. The same principle this whole series has been arguing since its first page — invalid states should be unrepresentable, not merely discouraged — applied, for the first time in this series, to what the humans in the loop are allowed to get away with.

> **Code and correspondence.** The work described here is implemented, not sketched. The source behind every claim in this paper — the engine, the extension, the fixtures and the tests that hold the invariants — can be shared on request, including the parts that did not work. Write to kenneth.pernyer@reflective.se. We are equally interested in being told where the argument is wrong.

References

Einarsen, S., Aasland, M.S., Skogstad, A., 2007. Destructive leadership behaviour: A definition and conceptual model. The Leadership Quarterly 18, 207–216.

Janis, I.L., 1972. Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes. Houghton Mifflin.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus, Giroux.

Pernyér, K., 2025a. What to learn before you train: formulating an Initiative Trajectory Model, and why training waits. Reflective Labs technical report.

Pernyér, K., 2025b. Institutionalized disagreement: how a plan earns the right to be run. Reflective Labs technical report.

Shaw, S.D., Nave, G., 2026. Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender. SSRN working paper.

Stasser, G., Titus, W., 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. Journal of Personality and Social Psychology 48, 1467–1478.

Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N., Malone, T.W., 2010. Evidence for a collective intelligence factor in the performance of human groups. Science 330, 686–688.

A. The failure modes, operationally

F1 Slop A proposal whose fluency exceeds any trace of the judgment that checked it. Detectable only as an absence of engagement in the process — missing counter-questions, no elapsed deliberation time, a pattern of first-answer acceptance — never as a trait of the approver.

F2 Hidden profile Evidence one participant holds that would change a group's conclusion, never proposed into the group's shared consideration. Distinct from being silenced: the participant was heard on other points and simply never said this one.

F3 Groupthink Suppression of dissent under pressure to maintain agreement, producing high collective confidence independent of whether the underlying reasoning was actually tested.

F4 HiPPO Convergence on the view of the highest-status participant, independent of the evidence, measurable as a correlation between a contribution's standing in the record and the contributor's rank rather than the contribution's content.

F5 Destructive leadership pattern Systematic, repeated conduct — tyrannical, derailed, or supportive-disloyal — that undermines the organization's own goals or its people's wellbeing. Assertable only against a documented incident record, never a single interaction.

B. Where the substrate's guarantees stop

Let S be the Suggestor fleet and $\to$ the step relation of my first report. Starvation freedom (axiom A7) states: every Suggestor eligible at cycle n is woken at some cycle $m \geq n$ before convergence is declared.

This is a guarantee about **scheduling** — that an eligible participant is not starved of the opportunity to propose. It says nothing about the **content** of what gets proposed once a participant is woken. Formally, for a human participant modelled as a Suggestor h, starvation freedom guarantees h is invoked; it does not constrain $\delta_h(K)$, the set of proposals h actually emits, beyond whatever h chooses to propose. A hidden profile is exactly the case $f \notin \delta_h(K)$ for some fact f that h holds and that would, if proposed and promoted, change the Formation's fixed point — a case the invariant is silent on by construction, because the invariant was built to guarantee execution, and disclosure is not a property of execution. Closing this gap requires a different kind of invariant, one that constrains δ_h rather than merely guaranteeing h runs, and this paper does not propose one.

