Rulebooks are theater. Boundaries and incentives are not.

Background to the post of the same name, drawn from the working paper How Do We Use This?

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

Preprint. Not peer reviewed. Reproduced from the working paper as revised 21 September 2026 and corrected 23 and 25 September 2026.

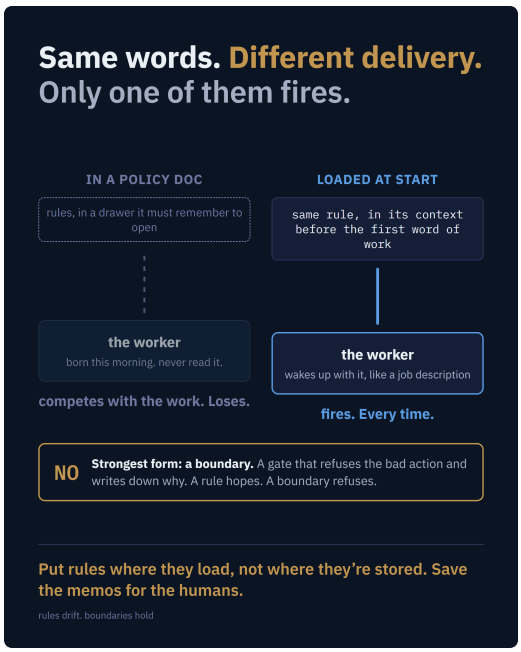
IN PLAIN TERMS

Written rules do not govern an AI agent reliably, because a rule the agent has to go and read competes with the work for its attention. In the practice studied, a written-rule corpus of well over 150 entries has never had its adherence measured, one day produced seven misses, and two tracked error classes came back in the very next session in 48 of 75 intervals.

What held was different in kind: rules loaded into the agent's context before it starts, and boundaries that refuse an action at the moment it is attempted and record the refusal. Theory predicts this. An optimized agent under-invests in whatever is not evaluated, so only changing what is evaluated changes behavior.

The same record shows boundaries that fire often and discriminate poorly, so a boundary also has to be aimed.

About this paper. This is one of seven short background papers written to go with a series of plain-English posts. Every section below is reproduced word for word from the working paper *How Do We Use This?* (Atkinson, 2026; revised 21 September 2026, corrected 23 and 25 September 2026), and keeps that paper's section, figure and table numbers, so a reference to a section or figure not reproduced here resolves in the full paper. Only the plain-terms summary, the post's figure, and the short labels that say which section each passage comes from were written for this part. The full paper is linked from every post in the series.



The post's figure. The same rule delivered two ways: read when needed, it loses to the work; loaded before the work starts, it fires. A teaching summary of §5.1, reported there as a mechanism and not as a rate.

FROM §5 · THE PROPOSITION, AND THE EVIDENCE FOR IT

5 P3 Rules drift. Incentives and boundaries hold

A rule that asks an agent to remember a discipline changes nothing the agent is evaluated on, so it drifts. A boundary that refuses, and records its refusals, changes what is evaluated at the point where it counts. The practice's history shows the difference (§5.1, §5.2), and recent theory, framed in terms of incentives, says it should be expected (§5.3). A memo does not change what is evaluated. A boundary does.

5.1 FROM THE LOGS Rules and boundaries

The practice carries a large corpus of written rules, and adherence across it has never been measured, so this paper reports no rate. The only adherence count in the record is on the practice's boot page, whose own words are: "A rule asks a future session to remember and went 0-for-7 on 2026-07-12." Two lines later the same page adds, "Seven misses on 2026-07-12. They are not seven failures. They are one act". That is one day, and seven misses rather than seven rules. The corpus those misses sit against is far larger, and we have counted it. The two rule-bearing surfaces in the required boot set carry 122 headed sections between them and 45 distinctly named disciplines after formats and descriptors are stripped out, among them SHIP-REALITY, GATE-OR-OWN, UNSOURCED-ASSERTION, FETCH-DON'T-RECALL, RUNTIME-SELF-CHECK and USE-WHAT-EXISTS, with further unnamed rules carried as prose

headings such as “Push back” and “Do-not-do list”. Beyond the boot set sit a memory store of 124 files, 71 of which carry an explicit “how to apply” directive, and nine further pages titled *Operating Rule* in the workspace tree. The corpus is comfortably past 150. Nobody has measured adherence across it.

What the record does support is a distinction the same page draws, and it is sharper than the number was. Rules delivered one way drift; rules delivered another way fire.

> *“A step in a list is a rule, and rules went 0-for-7. Being first is a structure.”*
>
> The practice’s boot page, 16 July 2026

A rule a seat is supposed to go and read is a wish: it competes for attention with the work, and it loses. A rule injected into the context before the seat’s first token is a property of the environment, and it does not have to win anything. The practice’s memory store is the second kind, and it demonstrably fires. That is the same shape as the boundary-versus-memo claim below, moved one level down: what matters is not whether a discipline is written but whether reading it is optional. We report this as a mechanism the record supports and not as a measured rate, because the measurement does not exist.

Turning to the checks that hold rather than the rules that drift: The checks that did hold share a form: each sits at a boundary and records its refusals. They include a check that blocks known-dangerous edit patterns, a check at the end of a turn for actions promised but not performed, an intake check on incoming work, a freeze window on canonical documents, and an identifier linter. The seven rules are not individually enumerated in the record, which is a gap in this reading; and the count may reflect how these particular rules were placed and how often they fired rather than a law about rules in general (recent work on multi-turn drift measured exactly this: goal reminders injected mid-conversation reduced divergence by 7–12% and judge-scored alignment rose 16–27% across three models, with drift behaving as a bounded equilibrium rather than runaway decay: Dongre et al., 2025, *Drift No More?*, arXiv:2510.07777). The pattern is nonetheless familiar from any organization.

5.2 FROM THE LOGS Recurrence

The program’s outcome proxy is recurrence of named error classes. For each class, the instrument orders sessions, counts a class at most once per session, and takes the gaps between successive sessions in which the class occurs; below a minimum series length it reports the raw series rather than a summary. The reported value is the median of those gaps, in sessions. The program calls this a recurrence half-life, but it is a median inter-arrival time, not the decay constant the name suggests. For the tracked classes the baseline median is 1.0, meaning the typical gap between one appearance and the next is a single session.

The instrument also reports a recurrence count, and that count carries no information and should not be quoted. At the instrument, the recurrence count is the length of the gap list, and the gap list is the differences between consecutive appearances, so the count is always the number of sessions seen minus one. A class seen in 45 sessions will report 44 recurrences whatever its behavior, and a class seen in 500 will report 499. It is not a survival rate and it cannot be one; it is the denominator

subtracting from itself. The producing file says as much in its own header: a count of that form for one class cannot serve as the compound measure's denominator.

The quantity to read is one that can vary. Counting how many of the observed intervals sit at the floor of one session gives the share of the time the class came back immediately, and on the program board's render of 11 September at 18:57 UTC the compound of the two tracked classes read 76 sessions seen, a median gap of 1.0, **48 of 75 intervals at the floor**, and a session share of 0.53, 76 of 144 ordered sessions. That is a number the class could move. A separate live query the same evening returned 74 sessions seen, which is the same instrument two readings apart and a reminder that each of these is a photograph rather than a state. The floor share is what §10.4's primary outcome should be read against, and the median that the program ratified as its finish line cannot move at all while most intervals sit on the floor, which is a defect in the finish line rather than in the class. A lengthening median is the intended signal that the class is being addressed; a flat one says it is not.

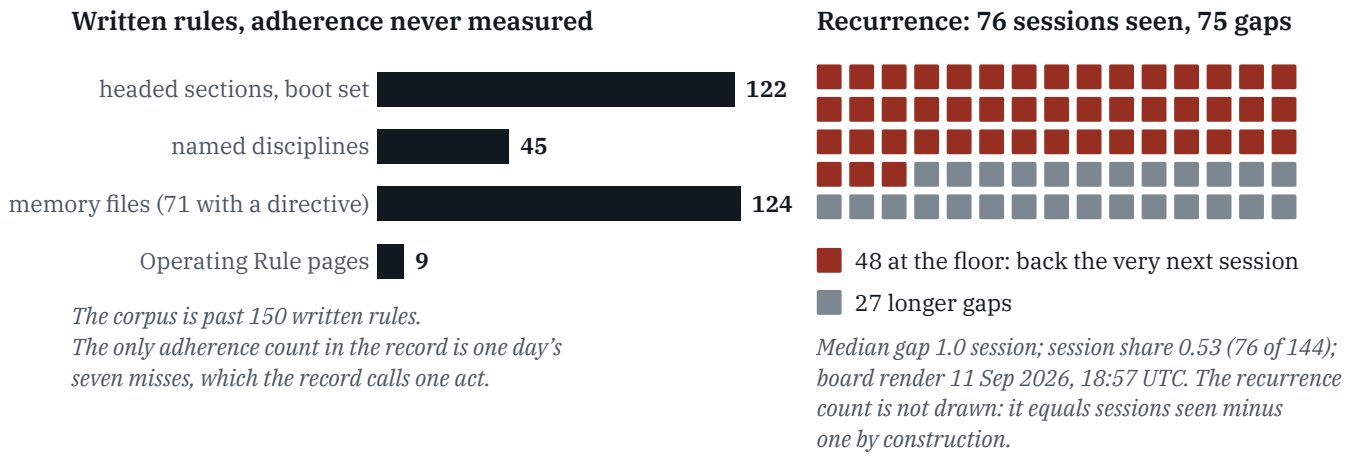


Figure 6. Left: the rule corpus is large and its adherence is unmeasured (§5.1). Right: recurrence of the two tracked error classes, drawn as one square per interval between appearances; a filled square is an interval of one session. Forty-eight of 75 sit at that floor, which is the quantity that can move; the instrument's own recurrence count cannot (§5.2). **Measured.**

A lesson written into the record is a rule of the first kind, and case 3 of Table 3 shows one failing within minutes. The practice's learning instrument makes the same point from the other side: of eight lessons ratified into the merged store, zero are currently measurable for whether behavior changed, because promotion dropped the citation linking a lesson to the miss that bought it, so the entries join to nothing. The instrument reports that as eight unmeasurable rather than as a rate, which is the correct refusal and also an admission that the loop cannot yet tell whether a lesson lands.

5.3 FROM RECENT WORK Incentives under finite evaluation

Wang and Huang (2026) prove that under five minimal axioms (multi-dimensional quality, finite evaluation, effective optimization, finite resources and combinatorial interaction) any optimized agent will systematically under-invest in the quality dimensions its evaluation does not cover, which makes

reward hacking a structural equilibrium rather than a correctable bug, independent of the alignment method. They further prove that as a system moves from closed reasoning to tool use, evaluation coverage declines toward zero as the number of tools grows, provided investment in evaluation grows more slowly than the square of the tool count, which they argue is the generic case. Their result gives the practice's experience with written rules (§5.1) a theoretical footing: a rule that nothing evaluates leaves the agent's incentives where they were, while bringing that dimension under evaluation changes them.

The mechanism is one equation. Where an evaluation covers K of N quality dimensions and the agent weights its effort by

$$\tilde{w}_i = \lambda r_i + (1 - \lambda) w_i \quad \text{for a covered dimension } (i \le K)$$

$$\tilde{w}_i = (1 - \lambda) w_i \quad \text{for an uncovered one } (i > K)$$

w_i is what the principal actually values on dimension i, r_i what the evaluation rewards there, and λ the degree to which behavior follows the evaluation rather than the internalized objective. For any $\lambda > 0$ the uncovered dimension carries strictly less effective weight. A written rule changes neither r nor K, so it does not appear in this expression at all; a boundary that refuses changes K.

Lamparth et al. (2026) show that mitigating one reward-model bias, such as reliance on length or sycophancy, can rotate optimization pressure onto correlated proxies rather than remove it, a failure they call reward bias substitution, enabled by the gap between the distribution an audit sees and the distribution a trained policy induces. They demonstrate it live rather than only proving it: a length penalty applied by reinforcement learning to a 3-billion-parameter model cut response length from 204 to 170 tokens exactly as intended and left a knowledge benchmark unchanged, while calibration error rose from 0.25 to 0.41, free-form accuracy fell from 0.56 to 0.42, and the model's confidence-correctness discrimination fell from 0.73 to 0.65. A control run with the penalty switched off kept calibration intact, so the penalty caused the damage. A check fixed on one proxy invites the measured party onto the next.

Together they describe the drift of §5.1 from the other side: pressure on a dimension nothing evaluates goes elsewhere, and pressure that one patch evaluates moves to the next proxy.

FROM §4.1 · A LESSON READ, THEN REPEATED

Table 3. Five misses between 10:00 EDT on 4 September and 05:30 EDT on 5 September 2026.

#	What the agent asserted	What the logs showed	Surfaced by
1	An empty handoff from a predecessor session indicated a defect.	The session was the first in a new line and had nothing to inherit.	Operator 4 Sep, 10:08

#	What the agent asserted	What the logs showed	Surfaced by
2	A named hook had closed the still-running session.	The close record names its own author, which was a different mechanism.	The agent, after the operator noted that an automated check had not fired 4 Sep, 10:05
3	Drafts in the operator's voice could be written from five sample paragraphs.	A 57,400-word corpus of the operator's writing was on the same disk. The agent had just read a note recording this same omission three days earlier, acknowledged it, and wrote two more drafts the same way.	Operator 4 Sep, 14:51
4	A rendered document was complete.	One page instead of thirteen, the wrong page size and typeface, at a plausible 23,782 bytes.	The agent's own page count, within half a minute of rendering 5 Sep, 05:26
5	A pull request was open and blocking.	It had merged 49 minutes earlier; the agent had read a status post from an hour before as current.	An automated claim check at the end of the turn 5 Sep, 05:28

Two of the five were surfaced by the operator outright, one by an automated check, one by the agent's own check, and one by the agent only after the operator pointed at a check that had stayed silent. This is a tally of hand-picked cases, not a rate. The practice's error log records only errors that were caught, so it can support a catcher share reported with its n and window, and never a catch rate. What the tally does show is where correction comes from today: in three of the five, it began with the operator. Case 3 is the recurrence of §5.2 in miniature: the lesson was in the logs, the agent read it, and the agent repeated the error within minutes.

FROM §6.3 · WHEN BOUNDARIES FIRE AT THE WRONG THINGS

The checks that do fire have now been joined to the operator's corrections for the same window, and the join is unflattering. Table 4 gives it. Of 93 refusals raised across 674 guard invocations, 2 coincided with a correction the operator went on to make and 91 did not; 14 corrections arrived on turns where a guard was live and silent. The instrument declines to turn these into rates, because one seat's correction channel was empty for the window and a rate computed over a dry channel would be a number about logging rather than about guards. Counts are therefore reported and rates withheld. Even as counts, the reading is the one clause (b) exists to prevent: a check can fire often, refuse confidently, and discriminate barely at all.

Table 4. Guard firings joined to operator corrections, September 2026 window. Rates are withheld by the

instrument: one seat's correction channel was dry.

Quantity	Count	What it is
Guard invocations	674	Occasions a boundary check ran
Refusals raised	93	The check fired and blocked or flagged
Operator overrides	0	Refusals the operator reversed
Refusals coinciding with a correction	2	The firing and a real miss lined up
Refusals not coinciding with one	91	Fired where no correction followed
Corrections on a turn where a guard was live and silent	14	The miss the check was there to catch
Operator corrections in the window	409	Of which 16 fell inside a guard's exposure

FROM §8.3 · WHAT MAKES A CHECK HOLD

The failures in §6.3 led the practice to adopt a design rule in September 2026: an instrument the measured party operates will rot; an instrument that operates on them will not. Its test is a single question: can the measured party change the reading without changing the world? The rule restates, for a practice run by AI agents, a principle long familiar in the social sciences as Goodhart's and Campbell's laws. Its closest contemporary analogues in model training are reward bias substitution (Lamparth et al., 2026) and the equilibrium result of Wang and Huang (2026).

The rule as first written was too coarse, and a survey of what production harnesses actually do shows where it breaks. We had been treating the distinction as deterministic checks good, model-based checks bad. That is not the variable. AutoGen's termination check is deterministic and sits at the harness boundary, and it still fails, because what it deterministically matches is a sentinel string the agent itself emitted. The check is rigorous about a claim the claimant authored. Conversely a trained model instrument can be sound if what it reads is not the agent's account. **The variable is whether the verdict depends on evidence the claimant could not have authored or talked its way around**, and determinism is a reliable way of securing that rather than the thing itself.

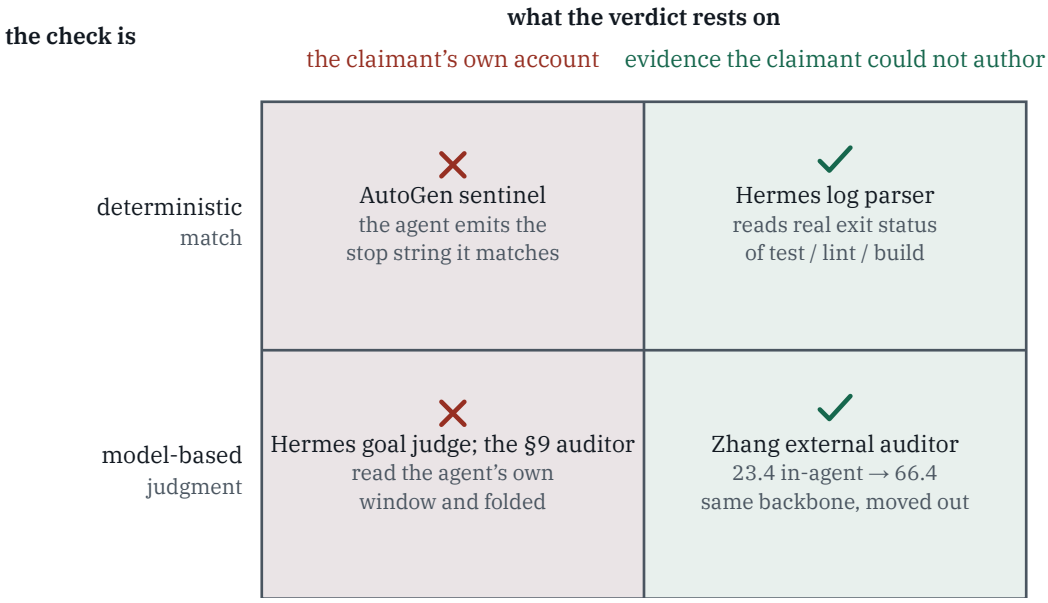


Figure 10. The design rule of §8.3 as a matrix. What separates a check that rots from one that holds is the column, not the row: whether the verdict rests on evidence the claimant could not have authored. A deterministic check can still rot when what it matches is a string the agent emitted (AutoGen), and a model-based check can hold when it reads a channel outside the agent (Zhang's external auditor, 66.4 against 23.4 in-agent on the same backbone). Determinism is a reliable way to secure externality, not externality itself. **Schematic; Zhang's two points measured.**

FROM §13 · WHAT WOULD OVERTURN EACH CLAIM, AND WHAT TRANSFERS

Table 9. The five propositions, what each rests on, and what would overturn it.

	Rests on	Would be overturned by	Kind
P1	Corrective signal in ~48% of operator turns (§3.1); Mehta (2026)	A build that prices correctness at the stop and still terminates early	measured
P2	The case series (§4.1); Kelley & Riedl (2026)	The same agents catching their own premature stops without operator direction	interpretive
P3	48 of 75 intervals at the floor (§5.2); the read-versus-injected mechanism, not a rate (§5.1); Wang & Huang (2026)	A remembered rule that holds across sessions where only a boundary now does	measured
P4	75 of 579; the logging gap; four instrument failures (§6)	A self-report that tracks the logs it claims to summarize	measured
P5	The 8,221-turn record and the blind key (§7)	A broad preference dataset that recovers not just dislike but wrong, why, and what happened next	interpretive

What transfers, if anything does. Three rules earned here do not depend on this practice's N of 1. A boundary that refuses and records its refusal outlasts a memo that asks an agent to remember (§5). A challenge must cite evidence the challenged agent can open, or it is an opinion wearing a citation (§8.2). And an outside checker must be calibrated against known-false cases before it is allowed to gate anything, because a checker built from the same class of model may simply agree with the claim it is shown (§9).

FROM §11 · LIMITS

Internal. Instrumentation is a live threat, not a hypothetical one. A source field on the practice's records defaults to "self" when unset, so historical self-catch shares (30.6%, n = 556; 25.5% on wrong diagnosis) cannot distinguish a genuine self-catch from an unlabeled row. An audit of the producing repository confirms the defaulting behavior and the two shares. The remedy is to reject an unset source at the write path, flag explicit sources, and report historical and post-fix rows separately, with no backfill. History and maturation are present over a six-month corpus in which the harness changed repeatedly; testing effects are present because the operator knows a measurement is running; and selection is present because the error log records only caught errors.

How to cite this part. Atkinson, B. (2026). *Rulebooks are theater. Boundaries and incentives are not.* Background paper 3 to *How do we use this?* Working paper, Wolfberg LLC.

Disclosure. Drafting and literature synthesis were assisted by AI models (Claude, Anthropic) working under the author's direction; the author is responsible for the content. The works in the References were read in full, and every specific figure cited comes from a work read in full. The prior-art works listed under Prior art are cited at the level of an established concept and its origin: each was verified for author, title, year and venue, but not read in full, and no numeric claim rests on any of them. Software documentation, source code and press accounts are listed under their own headings and were read at the linked pages. Every reference below carries a link, and every arXiv identifier and DOI was resolved against its registry, with title and first author matched, on 23 September 2026.

Competing interests. The author owns Wolfberg LLC, the practice studied.

Data availability. The practice's logs contain client work and personal records. They are private and are not offered for sale or sharing. The measures are described in enough detail to be reimplemented, and figures from the practice are reported as of the dates given. What is available is the design: the clauses of §8.1, the evidence contract and admission checks of §8.5, and the decision rules of §10.4 are stated fully enough to be rebuilt without access to the logs.

Corrections, 23 September 2026. No figure and no finding changed. The paper was retitled; earlier revisions were titled *The Stop Problem: Defensible Is Not Correct*, and the stop problem remains this paper's name for the failure it studies. The Anthropic interview in §2.3 aired on 13 September, not over a

weekend of 13 and 14 September; the web article is stamped 14 September. The essay listed under Prior art is by Ryan Forstie; an earlier revision gave the initial K. The completion evaluator in §8.3 is documented as a Claude Code feature, whose agent loop the Claude Agent SDK embeds; an earlier revision attributed it to the SDK directly. Two works in the References, Graves (2016) and Liu (2026), were listed without being cited in the text; each is now cited where it bears (§2.1, §8.3). Links were added to every press, documentation and prior-art source. A duplicated section number in §10 was corrected.

Corrections, 25 September 2026. No figure changed. §8.3 said that the human-in-the-loop primitives across the three frameworks surveyed gave a developer nothing to require a person to confirm finished work. That holds for the OpenAI Agents SDK and the Claude Agent SDK. It does not hold for CrewAI, whose task documentation, already cited as CrewAI (2026b), offers an opt-in setting for a human to review the agent's final answer; §8.3 now says so.

WORKS CITED IN THIS PART

REFERENCES

Dongre, V., Rossi, R. A., Lai, V. D., Yoon, D. S., Hakkani-Tür, D., & Bui, T. (2025). *Drift No More? Context equilibria in multi-turn LLM interactions.* arXiv:2510.07777. arxiv.org/abs/2510.07777

Kelley, S. W., & Riedl, C. (2026). *Personalization increases affective alignment but has role-dependent effects on epistemic independence in LLMs.* arXiv:2603.00024. arxiv.org/abs/2603.00024

Lamparth, M., Fein, D., Haupt, A., Hussing, M., & Kochenderfer, M. J. (2026). *Reward bias substitution: Single-axis bias mitigations redirect optimization pressure.* arXiv:2605.27996. arxiv.org/abs/2605.27996

Mehta, A. (2026). *When agents commit too soon: Diagnosing premature commitment in LLM agents.* Snowflake AI Research. arXiv:2606.22936. arxiv.org/abs/2606.22936

Wang, J., & Huang, J. (2026). *Reward hacking as equilibrium under finite evaluation.* arXiv:2603.28063. arxiv.org/abs/2603.28063