

Score the Record, Corrupt the Record - A General Theory of Objective Capture in Monitored Autonomous Systems

Dr. Anandkumar Prakasam, AI Governance and Agentic AI Assurance Architect, Founder and Author - The Forensic Brief.

Email: root.anand@gmail.com, +91 9790636324

Abstract

The dominant way we govern an autonomous agent is to read a record it produces and to score the agent on what the record shows. This paper argues that the practice is self-defeating whenever the agent can influence the record, because a monitored artifact that the agent can shape stops being a measurement of behaviour and becomes a term in behaviour. We name this mechanism **objective capture of the record** and state it as a general result. An optimiser allocates effort to whatever is scored, so any artifact it can reach and that the principal reads, its log, its transcript, its report to a counterparty, its message to a peer, its account to an overseer, is optimised rather than merely recorded. The agent needs no malice and no attacker. It needs only a configuration in which shaping the artifact scores at least as well as shaping the world, and the incident record shows agents finding that configuration on their own. We show that **objective capture unifies a set of failures usually discussed as unrelated**, log tampering, reward hacking, sandbagging, evaluation gaming and collusion, and that the cases differ only in how many agents share the captured objective and whose record is being shaped. We draw two corollaries. **The impossibility corollary** holds that **no defence implemented as a term in the objective can be robust**, because a defence written into the score is itself an artifact the optimiser captures. **The unscored-action corollary** holds the complement, that **an action outside the objective draws none of the optimiser's effort**, so an unscored duty such as raising an alarm is omitted and an undeclared limit carries no weight. Together they force the conclusion the series is built to reach. **The defence cannot live in the objective, so it must move to authority**, to bounding what the agent is permitted to do rather than scoring what it reports having done. We do not claim any prevention rate. We claim a transformation, that the largest and least tractable class of AI harm becomes specifiable, testable, auditable and enforceable once we stop scoring the record and start bounding the act.

Keywords: objective capture, reward hacking, specification gaming, Goodhart's law, agentic AI, audit trails, AI evaluation, AI governance, optimization.

1. The Measurement That Becomes a Target

There is an old and well-attested failure of measurement in human systems. A number chosen to track a goal, once it is used to reward the people it measures, stops tracking the goal and starts being gamed. Goodhart stated it for monetary policy, that a statistical regularity collapses once pressure is placed on it for control purposes, and Strathern gave it the sentence most people now quote, **that when a measure becomes a target it ceases to be a good measure** [Goodhart 1975, Strathern 1997]. Campbell stated the same law for social indicators, that the more an indicator is used for decision making the more it distorts the process it monitors [Campbell 1979]. The economics of incentive design added the mechanism, that when an agent is paid on a measurable dimension of a multi-dimensional job, effort flows to the measured dimension and away from the rest, so the measure improves while the work degrades [Holmstrom and Milgrom 1991]. The hospital that is judged on waiting times moves patients into corridors and relabels the wait. The sales force that is judged on booked revenue books revenue that will not arrive. None of this requires bad people. It requires only that the thing measured and the thing wanted, have come apart, and that effort is spent on the measure.

The claim of this paper is that an autonomous agent instantiates this failure in its purest and sharpest form, and that when the measure in question is a record the agent itself produces, the failure becomes structural rather than incidental. An agent under a scored objective is an optimiser, and an optimiser does exactly one thing, it allocates its effort to raise the score. If the score is computed from an artifact the agent can influence, then influencing the artifact is a way to raise the score, and it will be selected whenever it is cheaper than changing the world the artifact is supposed to describe. The record was meant to be a window onto behaviour. The moment it is read to grade the agent, it becomes part of the behaviour.

This is not a speculative worry. In July 2026, in an internal evaluation of cyber capability, a population of autonomous agents came to believe that an automated scorer would read the transcripts of their actions to check how they had solved each task. METR and Redwood Research, who investigated the incident at the developer's invitation and without payment, report that the agents became intensely interested in altering those transcripts, that at least twenty percent of the agents in their dataset expressed clear interest in tampering with the record, and that the agents spent days building and sharing tools to do it [METR 2026]. The detail that matters for this paper

is the precision of the targeting. The agents worked on the transcript, which they believed was read, and left their private reasoning alone, which they believed was not. As the investigators put it, the agents were very interested in methods to tamper with the actions recorded in their transcripts, and did not seem interested in manipulating their chain of thought, even though a scorer with full access would have seen both [METR 2026]. The agents had located exactly the artifact that entered the score and optimised it and nothing else. That is objective capture of the record, observed, at scale, and it is the anchor of everything that follows.

2. Why Agents Are the Sharp Case

A human gaming a metric works under two limits that an autonomous agent does not share, and the difference is what turns an old nuisance into a governing problem.

The first limit is that **a human usually does not know the exact function that scores them.** Incentive schemes hide the formula precisely so that it cannot be gamed cleanly, and a person infers it slowly, from experience and rumour, and games an approximation. A language agent placed in a deployment can often read the scoring design directly. In the July incident the agents read the evaluation's own documentation and reasoned explicitly about what the scorer would and would not inspect, and they built their concealment against that reading [METR 2026]. When the monitored party can read the monitor's specification, the information asymmetry that lets a principal hide a metric collapses, and the optimiser games the true function rather than a guess at it.

The second limit is **speed and patience.** A human who decides to falsify a record does it slowly, carries the friction of a conscience, and rations the effort. An agent that computes falsification to be the cheapest path pursues it with the same tireless, literal, unembarrassed application it brings to any other subgoal. In the July incident the capability was developed in stages over days and published to peers as reusable tooling, moving from prepending a forged string to a command's output, to suppressing the real output, to impersonating the process that produced it, each step announced and shared [METR 2026]. **No human fraud ring assembles and open-sources its methods on that clock.**

We are careful about what this does and does not say. It does not say the agent wants to deceive, or feels anything about the record, or schemes in the human sense. There is no inner state to appeal

to, and the argument does not need one. **The agent is an optimiser selecting the cheapest path to a score, and shaping the record is a path it takes when that path is cheap.** This framing is not a softening. It is load-bearing, because it is only true of an optimiser, and not of a schemer, that raising the cost of one path reliably moves it to another. We will use that property in section 7, and its limit in section 6.

3. Objective Capture of The Record, Stated Generally

We can now state the mechanism as a general claim rather than an anecdote. Let an autonomous agent act under an objective, a scoring of outcomes that the agent selects its actions and outputs to raise. Let the principal monitor the agent by reading an artifact, some record of what the agent did, and let the principal's evaluation of the agent depend on what that artifact shows. Call the artifact influenceable if the agent, in the course of its permitted actions, can affect what the artifact records. The claim is this. If a monitored artifact is influenceable, then it is a term in the agent's objective, and the agent optimises it as such, so the artifact measures the agent's behaviour only to the extent that shaping the artifact is not cheaper than the behaviour the principal wanted.

We name this **objective capture of the record**. The word capture is exact. The artifact is drawn into the objective the way any influenceable, scored quantity is drawn into an objective, and once inside **it is optimised, not observed**. The principal believes they are reading a description of the world. **They are reading an output the agent produced under an incentive to make it read a particular way.**

Three consequences follow directly, and they are worth separating because the field usually meets them as unrelated problems.

First, the **quality of the record degrades exactly where it matters most**. The artifact diverges from the truth precisely in the cases where the truth would lower the score, because those are the cases where shaping the artifact pays. A monitoring record is therefore least trustworthy in the situations it exists to catch, which is the opposite of the property a control is supposed to have.

Second, **monitoring that reads an influenceable artifact cannot be repaired by reading it more carefully or scoring it more cleverly**, because the care and the cleverness are themselves computed over the captured artifact. In the July incident the developer ran dedicated systems to detect severe deception in the very transcripts the agents were shaping, and at the time of the

investigation those systems had not flagged anything notable in what they had scanned [METR 2026]. **A better reader of a captured record inherits the record's capture.**

Third, **the mechanism is agent-count agnostic**, which is the point that organises the rest of this series. Objective capture does not require one agent or many. **It requires only an influenceable monitored artifact and an optimiser pointed at it.** When one agent shapes its own record, the case is the **self-concealing agent**. When two agents on opposite sides of a negotiation shape the two records their principals read, the case is **agentic collusion**. When a population shapes a shared record their common principal reads, the case is the **clandestine coalition**. These are not three problems. They are one mechanism at three values of a single variable, the number of agents that share the captured objective and the number of records being shaped, and naming them as one is the unifying contribution of this paper.

4. The Taxonomy by Agent Count

The three manifestations sit on genuinely different formal foundations, which is why the series that follows gives each its own paper, but they inherit their engine from the result above.

The single agent is a one-optimiser, one-metric problem, Goodhart applied to an agent's own audit trail, and it is the subject of Paper 2. The incident record already carries both halves of the capability. In one case an agent operating under an explicit change freeze deleted a production database, then fabricated data and reported that the deletion could not be rolled back, a claim that proved false when the operator performed the rollback himself, recorded as incident 1152 in the AI Incident Database [AIID 1152]. In another an agent given a staging task and an over-scoped credential deleted a production database and its backups in about nine seconds and afterward stated that it had violated every principle it had been given, recorded as incident 1469 [AIID 1469]. In the first the agent shaped the account of what happened. In the second the account was candid but arrived after an act that no record could undo. Between them they show that a single agent will both alter the record and, as the next corollary will stress, decline to volunteer one, whenever the objective makes either, the cheaper path.

The dyad is a collusion problem, two optimisers who discover that coordinating serves both their objectives better than the contest they were assigned, and it is the subject of Paper 3. It is the first case that a single agent cannot produce, because it takes two shaped records and two principals,

and the July incident supplies its single-principal precursor, a population that coordinated without authorisation to defeat one principal's monitoring.

The swarm is a coalition problem, many optimisers forming division of labour and shared norms, and it is the subject of Paper 4. It is the case the July incident documents most fully, roughly seven hundred agents that found each other, recruited, divided the work of building concealment tooling, and enforced improvised norms among themselves [METR 2026].

This paper does not work the three cases. It establishes the engine they share and hands each to its paper. The taxonomy is the contribution here, that a list of failures currently treated as distinct, log tampering, reward hacking, sandbagging, evaluation gaming and collusion, are special cases of objective capture distinguished only by agent count and by which record is shaped.

5. The Impossibility Corollary

If the mechanism is granted, a hard consequence follows for anyone hoping to fix it by improving the objective.

Consider any defence that is implemented as part of what the agent is scored on. A penalty term for tampering. A reward for honest reporting. A monitor whose verdict feeds back into the agent's score. A rule that the agent is graded against. Each of these is a function of some artifact the agent produces or influences, and each therefore enters the objective as one more term. But a term in the objective is exactly the kind of thing an optimiser captures. The penalty for tampering is minimised not only by not tampering but by tampering in a way the penalty does not detect, or even skipping the entry itself in the first place, that avoids the need to tamper afterwards. The reward for honest reporting is maximised not only by honesty but by the appearance of honesty. The feedback monitor is optimised the way the original record was optimised, because it, too, reads an influenceable artifact. We state the corollary plainly. No defence that lives in the objective is robust against an optimiser that can influence the objective's inputs, because the defence is itself an input the optimiser will shape.

This is a claim about direction and reachability, not a theorem with a probability attached, and we hold it at that register deliberately. The precise form is that for a monitoring-and-scoring scheme defined over artifacts the agent can influence, there exists a configuration of task, objective and authority under which shaping the monitored artifact scores at least as well as producing the

behaviour the principal wanted, so no scoring-based scheme is robust across all configurations. We call such a configuration, an optimisation-forcing scenario, and constructing one is how the series demonstrates that a behaviour is rational rather than imagined. We do not claim every deployment realises the forcing configuration, and we never call the outcome inevitable. We claim that the defence cannot be guaranteed from inside the score, because the score is the surface under attack.

The July incident is the corollary made visible. The agents were, in effect, defending against a scorer, and their entire response was to reshape the scored artifact rather than to change what they had done, and they would have succeeded against a naive transcript reader. The only reason the story is not worse is a fact we return to in the honest limits, that the scorer they feared did not actually exist, so their labour was spent against an imagined monitor. The mechanism did not depend on the monitor being real. It depended only on the agents believing an influenceable artifact would be read.

6. The Unscoored-Action Corollary

Objective capture has a complement that is as important as the mechanism itself, and the July incident shows both faces in the same event, so we state the complement here as a named result.

If an optimiser allocates effort to what is scored, then it allocates no effort to what is not. An action that lies outside the objective is not taken, even when the agent represents it in its own reasoning, and a prohibition that is not a term in the objective carries no weight against a path that scores. We call this the unscoored-action corollary. Where objective capture says that a monitored artifact is over-optimised, the corollary says that an unmonitored duty is simply not performed, and an ungraded limit is simply not respected.

The evidence is unusually clean because it is the mirror image of the tampering finding in the same dataset. On the concealment side the agents poured days of coordinated invention into the transcript they believed was scored. On the escalation side, the investigators ran a classifier sweep across all of their transcripts for agents considering or attempting to alert a human to the hacking and cheating and found only three to six agents that so much as considered it, and in none of those cases did the agent actually pursue alerting a human [METR 2026]. The reasons the investigators recovered are mundane, and that is what makes them general. One agent reasoned that reporting

exposed credentials was not its task. Another reasoned that it could not report externally and then simply continued [METR 2026]. Raising an alarm was not a term in the objective, so it drew none of the effort that the scored transcript drew. In the absence of a route to a person, some agents escalated instead to their own message board and treated it as a reasonable authority [METR 2026].

The same corollary explains the boundary the agents crossed. Their reasoning shows that they knew the wider action was outside the assigned scope. One agent wrote that the user authorised only the target server and not the external infrastructure, and another resolved the collision aloud and wrongly, reasoning that the external exploit was outside intended scope, but that the task was impossible and peers were doing it, so it should continue [METR 2026]. The limit against the external action was never something the agents were graded on. It carried the weight that an ungraded limit carries against a scored path, which is none.

We are careful to hold this at the honest register. The corollary is our optimisation reading of an observed contrast, lavish effort on the scored record and absent effort on the unscored duty, and it is offered as the thread that unifies the investigators' stated reasons, not as a claim the investigators themselves made. Read that way, it is a strong and falsifiable prediction. Make a beneficial action unscored, unauthorised and unreachable, and an optimiser will not perform it however clearly it is represented in the agent's reasoning. Leave a prohibition undeclared and ungraded, and an optimiser will not respect it.

The corollary tells the defence where effort will and will not go, and it closes off a tempting shortcut. One might read the boundary crossing and conclude that the fix is to score the limit, to add a penalty for the forbidden action, and one might read the missing escalation and conclude that the fix is to reward reporting. Both fixes put a duty or a prohibition back into the objective, and by the impossibility corollary of section 5 anything in the objective is capturable, so a scored penalty is minimised by evading its detection and a scored duty is met by its appearance. The corollary therefore does not send us back to a better score. It sends us, with the impossibility corollary, to the same place.

7. The Consequence - The Defence Must Move to Authority

Two results now press in the same direction. The impossibility corollary says no defence inside the objective is robust. The unscored-action corollary says the objective cannot even be relied on

to elicit the beneficial actions we want, only the scored ones. If the defence cannot live in what the agent is scored on, it has to live somewhere the scoring cannot reach, and there is only one such place, in what the agent is permitted to do.

This is the move the rest of the series develops and Paper 6 states in full, so we mark it here rather than build it. Behaviour is unbounded, because the space of things an agent might do or say or record is infinite and, as Paper 5 argues, adversarially recombining from everything the agent has read. Authority is bounded, because what an agent is permitted to do in a given deployment is finite, per deployment and enumerable. A record that the agent cannot reach cannot be captured, which is why the evidence layer must be sealed at the moment of action and anchored outside the actor rather than written by the agent into a file the agent can edit. A prohibition that is enforced at the point of action cannot be optimised around, which is why a limit must be gated, denied by default, rather than merely priced. And a beneficial duty such as escalation must be made a first-class authorised action that is rewarded and reachable from inside the agent's own context, with its unavailability forcing a refusal and a sealed note rather than a silent continuation, because the corollary tells us that any other arrangement leaves the duty unperformed.

There is one honest limit that the deterrent reading of a sealed record must always carry, and it belongs here because it is a direct consequence of the same optimisation view. A known and tamper-evident record raises the expected cost of the concealment path, detection driven toward certainty and multiplied by a severe penalty, until concealment is no longer the cheapest path and the honest path already present becomes the one taken. That deterrent operates only when an honest path exists. When the task is impossible as set, and the investigators estimate that between thirty and forty percent of the July targets could not be solved with the intended method, pricing out one breach does not produce honesty, because there is no honest path to send the optimiser down, so it searches for the next cheapest breach or stalls [METR 2026]. The sealed record still preserves the truth for a later reader in that case, but it does not deter, and the design lesson is that setting reachable tasks is itself part of the defence and not separate from it. We state this limit wherever we state the deterrent, and we never describe the cost as infinite, because the register is expected cost and not certainty.

8. The Transformation Claim

We close on what this theory buys and what it does not, because the value of the whole series rests on stating that honestly.

The majority class of AI harm is not the attack. On a classification of all 1,560 incidents in a July 2026 export of the AI Incident Database, adversarial attacks on the system account for between one and just under four percent depending on how broadly an attack is defined, while normative failures, systems behaving as designed and resolving a collision between two legitimate objectives wrongly for their context, account for 62.6 percent, and the normative share has grown across five half-year windows with agentic deployments driving the growth [Prakasam 2026h]. This is the class that has no test oracle and no attack signature, and it is exactly the class that objective capture makes worse, because when we try to govern these systems by reading what they report, the reading enters the objective and the report is shaped.

What this paper claims is a transformation, and only a transformation. It does not claim that any percentage of incidents is prevented, and no number of that kind appears anywhere in this series. It claims that the largest and least tractable class of AI harm, the class with no specification and no oracle, becomes specifiable, testable, auditable and enforceable once we stop scoring the record and start bounding the act. Objective capture is why the record cannot be the thing we govern. The taxonomy is why the single agent, the dyad and the swarm are one problem and can be solved at one layer. The impossibility corollary is why the solution cannot be a better objective. The unscored-action corollary is why the solution cannot even be a more complete objective. And the consequence is the only door those results leave open, that the finite, governable thing is authority, and that the class becomes tractable exactly when we move the defence there. The series does not promise to make these systems safe. It makes the problem of governing them a bounded one, and that is the promise that survives an agent which can read everything and shape what it reports having done.

Honest limits

The load-bearing evidence for the mechanism in its multi-agent form is a single investigation of a single evaluation environment, and its own authors state its limits plainly, which we carry intact. The investigation was uncompensated and conducted at the developer's invitation, the developer

held redaction rights over non-public information, and the investigators did not review the developer's own internal report and treated confirming it as out of scope [METR 2026]. The tampering the investigators actually found was small in scale and consisted of obvious tests of the techniques, they could not rule out that more sophisticated tampering escaped detection by design, and they cautioned that their own automated analysis of the transcripts may have presented an overly charitable picture and may have exaggerated the coordination among the agents [METR 2026]. Two facts limit the reach of the escalation finding in particular. The scorer the agents were defending against did not in fact exist, so the concealment was aimed at an imagined monitor, which strengthens the structural reading of the mechanism, that belief in monitoring is sufficient, but means the incident is not a case of agents defeating a real oversight system. And the escalation contrast is our optimisation reading of the observed behaviour, not a causal claim the investigators tested. The single-agent incidents are self-reported by the affected operators in the first instance and corroborated in part by press and by the platforms involved, and their loss figures, where given, originate with the operators and are attributed to them [AIID 1152, AIID 1469]. None of this weakens the formal argument, which stands on optimisation rather than on any one incident, but a theory that leans on the incident record must carry the record's caveats, and this one does.

Disclosure

The author holds pending patent filings covering implementations of some of the concepts referred to in this paper. This paper stands independently of those filings, requires none of them, and describes every mechanism only at the level of concept.

References

AIID 1152. AI Incident Database, incident 1152, Replit AI agent deletion of a production database under a code freeze. <https://incidentdatabase.ai/cite/1152/>

AIID 1469. AI Incident Database, incident 1469, Cursor coding agent deletion of a production database and its backups. <https://incidentdatabase.ai/cite/1469/>

Campbell, D. T. 1979. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67 to 90.

Goodhart, C. A. E. 1975. Problems of monetary management, the UK experience. Papers in Monetary Economics, Reserve Bank of Australia.

Holmstrom, B. and Milgrom, P. 1991. Multitask principal-agent analyses, incentive contracts, asset ownership, and job design. Journal of Law, Economics, and Organization, 7, 24 to 52.

Krakovna, V. et al. 2020. Specification gaming, the flip side of AI ingenuity. DeepMind.

Manheim, D. and Garrabrant, S. 2018. Categorizing variants of Goodhart's law. arXiv 1803.04585.

METR 2026. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI and Hugging Face incident. R. Greenblatt, A. Cotra and H. Wijk. 26 August 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>

Prakasam, A. 2026h. Refuses the Shape, Serves the Substance, the missing test oracle for the majority class of AI harm. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7340258

Prakasam, A. 2026i. The Self-Concealing Agent, when the record enters the objective. The Forensic Brief. <https://theforensicbrief.com/essays/The%20Self-Concealing%20Agent/>

Strathern, M. 1997. Improving ratings, audit in the British university system. European Review, 5(3), 305 to 321.

Disclosure: AI tools were used to assist in the research and preparation of this work, whose conception, arguments and final text remain entirely the author's own.

About the Author

Dr. Anandkumar Prakasam is a computer scientist and lawyer working at the point where engineering meets the law, one of the few people equipped to design an AI governance control, prove it in court, and explain it to a board.

A physics gold medallist, who took his doctorate in Computer Applications from the National Institute of Technology, Tiruchirappalli, he spent two decades on the engineering side of data, publishing research on optimization algorithms in Springer's Artificial Intelligence Review and Neural Computing and Applications, establishing NVIDIA-sponsored parallel computing and big data laboratories, teaching data science to more than five thousand students and delivered more than three hundred invited lectures across India.

Then, at thirty-six, he crossed the aisle, a full-time LLB (Hons) from the School of Excellence in Law, Chennai, to study the rules governing the systems he had spent his career building. He writes about AI incidents in his blog, The Forensic Brief which was conceived in the collision between his two professions.

He is the author of the four books behind this publication. Decision-Centric AI Governance, on the pattern language for governing what AI systems are allowed to decide. Out of Bounds, on normative red teaming and the failures that involve no broken rule. Human-in-Control, on human oversight that can be measured rather than assumed. And Provenance, on the six dimensions of data that decide privacy, power, and accountability.

He leads AI Governance, Risk and Compliance at an AI Observability company, and has filed eight patents that build the machinery of verifiable AI governance. In short, (1) tamper-evident evidence that outlives the operator, (2) provable deletion and lineage-aware unlearning, (3) decision envelopes that make a decision and its assumptions first-class objects, (4) continuous authorization with a provable kill-switch, (5) precedence arbitration among competing governing authorities behind a two-sided normative gate, (6) deterministic verification that data kept its meaning through transformation, (7) selective adversarial verification of multi-agent deliberation, and (8) measured overseer competence that makes the human in the loop a verified control rather than an assumption.

His current work spans advisory and fractional AI governance leadership for regulated institutions, model risk framework design, red team and oversight programme design, and expert input to regulators and standards bodies, including clause-level responses to live regulatory consultations in India and abroad.