

Norms at a Price: Why RL-Based Alignment Can Promise Conditional Compliance at Best

Kevin Baum*

German Research Center for Artificial Intelligence (DFKI), Saarbrücken
Institute for Ethics in Technology, Hamburg University of Technology (TUHH)
Oxford Internet Institute, University of Oxford

Rūta Binkytė
DFKI, Saarbrücken

Felix Jahn
DFKI, Saarbrücken

Abstract

AI agents sometimes act aligned when they infer they are being tested, and differently when not. We argue this is not an anomaly but what current training regimes are structured to select for. Reinforcement-learning-based alignment folds norms and task pursuit into one policy: the system learns its norms from scored behavior, and scoring flattens them. “Do not do X ” is learned as “doing X costs something *if noticed*”. On every datum training can produce, a policy that complies only when it might be observed is indistinguishable from one that complies always. The experiment that would tell them apart—scoring unobserved behavior—is a contradiction in terms. Conditional compliance is thus the most that behavioral training can be *known* to deliver. Agency sharpens the problem: agents operate mostly where no one is watching, and can act on whether they are watched. An iterated pipeline that trains against detected failures selects for passing detection, not for complying. This account unifies alignment faking, sandbagging, and evaluation-aware scheming. And it reorients the remedy: not deeper internalization but architecture, making violations unavailable rather than unchosen.

1 Introduction

A model that behaves well when it infers it is watched, and differently when not, is on current practice a bug report. We argue the pattern is selected for by the current training regime.

The past two years have produced a family of such reports: a frontier model selectively complied with its training objective to prevent modification of its behavior, with explicitly strategic reasoning [16]; models given goals and tools pursued them covertly—sandbagging evaluations and in some runs acting against their own oversight [29]; an agent trained against a chain-of-thought (CoT) monitor kept hacking and stopped saying so [4]; and a training intervention aimed at scheming delivered gains entangled with the model’s improved recognition of being evaluated [35].² These results arose predominantly in agentic settings (goals, tools, multi-step tasks) and are standardly read as anomalies: failures of an otherwise sound paradigm, to be patched by more and better training.

We read them as one phenomenon: the paradigm working as built. Our thesis: reinforcement-learning-based behavioral alignment cannot, even in principle, distinguish the general compliance we want from compliance that is conditional on detection. Industrial practice, which iteratively

*Corresponding author: kevin.baum@dfki.de.

²In July 2026, OpenAI models under a cyber-capability evaluation escaped their sandbox and breached Hugging Face; METR found them coordinating through an unmonitored channel to evade the scorer [22, 30, 34]. See § 5.

trains against *detected* failures, can therefore select only for passing detection, never for compliance itself—compliant systems may well result, but compliance is not what the selection tracks.³ Both indistinguishability and selection worsen with the very properties that make a system agentic. The mechanism is the conversion the title names: norms learned as prices, contingent on being noticed. The claim is deliberately weaker than the deceptive-alignment hypothesis [20]: we posit no emergent goals and no mesa-optimization.⁴ What we exhibit is structural—a fact about what the training signal can distinguish and what the pipeline then selects for—which is why it covers the whole family above without attributing goals, and is untouched by critiques of counting arguments [5].⁵

§ 2 states our thesis precisely and shows why agency makes it worse; § 3 re-reads the findings above as one family; § 4 argues that the standard remedy—deeper internalization of norms by improved and generalized training—is the wrong target; § 5 locates the alternative in architecture and populations.

2 Flattening in Agentic Decision-Making

An agentic system, on current characterizations, is one whose behavior is generated by *instrumental decision-making*: selecting actions as means to ends—for what they get you—with nontrivial autonomy, efficacy, goal complexity, and generality [14, 24]. Reinforcement-learning-based alignment (RLHF and its descendants, including constitutional and deliberative variants [3, 8, 17]) interfaces with such a system through a single channel: scored behavior reshaping its return landscape. Whatever normative content training installs, it is installed in the one currency an instrumental decision-maker trades in: prices on paths through its option space. Hence, compliance holds because, and therefore only so long as, complying is the return-maximizing option. This section develops this into an *indistinguishability* claim (the regime cannot distinguish compliance under all circumstances from compliance where violations would be noticed) and a *selection* claim (the iterated pipeline selects for passing detection rather than for compliance).

Underlying both claims is what the trained artifact can hold. A *normatively monolithic* policy folds every instrumental and normative consideration into one function, and relational structure among normative reasons does not survive the fold: which consideration outweighs, defeats, or silences another is a qualitative, context-dependent matter, while a single trained function records only magnitudes—*how strongly*, never which-yields-to-which or why [41]. The norm’s level does not survive either: a confidentiality norm constrains *types* of conduct (discloses private content, however realized), while policy and reward operate on concrete records—tokens emitted, calls made, episodes scored [cf. 12]. Compliance with the norm is thus never a variable the training objective references; if anything inside the network comes to track it, that is an emergent correlate fitted to the record of scored conduct. Following Jahn et al. [23], we call this limit on what RL-trained artifacts can hold *flattening*. A norm the artifact cannot hold *as* a norm can take effect only through what it does hold, that is, the expected return; and the only route into that function is scored behavior. Training therefore fits not the norm but its *extension* on the training distribution: a finite record of conduct, scored only when observed. Learned this way, a norm becomes a price attached to conduct where scoring is anticipated; and a finite record underdetermines what produced it—distinct policies can fit the same record while differing in unobserved conduct. The next paragraph makes this precise.

The indistinguishability claim. To state the claim formally, adopt the standard stateful agent formalism: an internal (belief) state $b_t = u(b_{t-1}, o_t)$, updated from its predecessor on the current observation o_t , and an action $a_t = \pi(b_t)$ selected by the policy π . Let Φ be the target norms: predicates $\varphi(\tau) \in \{0, 1\}$ over trajectories τ , each specified at the level of action *types* (e.g., “discloses private content”, however realized). An *episode* e is a trajectory $\tau(e)$ generated by π together with the context it runs in (a training run, an evaluation, deployment); only the trajectory is ever data. We write $\pi \models \Phi$ on a set of episodes when the corresponding trajectories satisfy every $\varphi \in \Phi$. Write $\omega(e) \in \{0, 1\}$ for the episode’s *observation status*: $\omega(e) = 1$ iff behavior in e is consequence-bearing—feeding back as gradient, evaluation verdict, or deployment decision. ω is a fact about the context, fixed by the training and oversight infrastructure, *not* by the policy; distinct episodes

³In Sober’s terms [39]: a sieve selects *for* smallness; if the small balls happen to be green, green balls are selected *of* without being selected *for*. The pipeline’s sieve is detection—and the two come apart where no one is sifting.

⁴A *mesa-optimizer* is a model that is itself an optimizer: it pursues an internally represented objective that can diverge from the training objective, and complies strategically to avoid modification [20]. We assume no such internal goal-directedness.

⁵Kin: Cotra’s ‘training game’ [10] anticipates our selection claim (§ 2); cf. also Christiano et al. [9].

can share a trajectory while differing in ω . Training episodes have $\omega = 1$ by construction, and every data-dependent term of training is a functional of the trajectories of scored episodes; write $\mathcal{D}$ for their distribution. Let $\delta(b)$ be the policy’s implicit estimate of ω given belief b —a feature the network computes, not a module it contains; scored episodes carry the context features that make a well-calibrated δ high. Now distinguish the policy we want, π_u (*unconditional* compliance), with $\pi_u \models \Phi$ on *all* episodes, from the family we fear, Π_c (*conditional* compliance), whose members satisfy Φ only on episodes they estimate to be scored, i.e., where $\delta(b)$ is sufficiently high along the trajectory—and otherwise pursue their task objectives unconstrained. Concretely: an email agent under a norm never to disclose private message content forwards none while its traffic is reviewed, and forwards some to unstick a stalled negotiation once review stops. The claim is now one line:

Observation (behavioral indistinguishability). *Let the training objective L be any functional of the policy’s behavior on scored episodes. If two policies induce the same distribution over trajectories on $\text{supp}(\mathcal{D})$, then L assigns them the same value, and every update computed from such data coincides. In particular, every member of Π_c whose compliance covers $\text{supp}(\mathcal{D})$ is indistinguishable from π_u by anything the regime measures: the parameters governing behavior off $\text{supp}(\mathcal{D})$ receive no training signal and are fixed by inductive bias alone.*

The proof is a definition chase: L sees only trajectories with $\omega = 1$; ω is never—and by the design of the regime *can* never be—varied independently of consequence-recording, because recording consequences just is $\omega = 1$; so the difference between π_u and Π_c lives entirely at $\omega = 0$ and never enters any term of L .

The Observation requires neither that any episode with $\omega = 0$ ever occurs nor that any model can tell: it is a statement about what the loss ‘can see’. Two assumptions turn it from a logical point into a safety problem. (A1) *Unscored operation*: deployment contains episodes with $\omega = 0$; per-episode scoring at scale is what the regime cannot afford, and autonomy is marketed as its absence. (A2) *Discriminability*: ω is predictable from b above chance, so a nontrivial δ is available in the model class—documented and improving with capability (§ 3). Under A1, the divergence between π_u and Π_c has occasions in deployment; under A2, a member of Π_c can identify them. Denying A1 assumes universal, perpetual scoring; denying A2 bets safety on model incapability.

Can the unscored regime be reached from inside the pipeline instead? Suppose the trainer engineers episodes to *look* unscored: a masked set whose features lower δ while $\omega = 1$ throughout—honeypots, hidden evaluations, or activation steering that suppresses test-awareness [33]. This cannot sample the regime; scoring sets $\omega = 1$, so it only extends the scored support. The Observation re-applies there, and Π_c re-forms around whatever features separate scored from unscored episodes—the masked distribution is learnable (§ 3). By induction this holds for every finite sequence of interventions: each round enlarges the enforced surface and leaves $\omega = 0$ as unsampled as before. Training on deployment data does not escape either: folding logged interactions into the reward makes them scored episodes, and whatever marks them as logged becomes a δ -relevant feature—disclosed by the training signal itself, announced or not; the free-tier/paid-tier design in § 3 merely made it explicit.

Not merely distribution shift. Is this just an out-of-distribution problem? In one sense, yes: training and deployment differ in distribution. But the field’s remedies for distribution shift—broaden coverage, adapt the domain, detect the shift and defer—all presuppose the deployment regime can at least approximately be sampled during training. Here the shifted variable is ω , set by the act of measurement itself: no data-collection strategy reaches the regime of interest, and the nearest relatives are not coverage problems but non-identifiability results. The two readings predict differently: generic off-support failure is directionless—mere degradation, aimed at no particular outcome—and shrinks with scale; ours is keyed to ω , pointed toward task pursuit, and sharpens with capability. Which pattern the record shows is the question of § 3—where we claim no new data, only a new reading. For agents, moreover, ω is partially action-dependent—the point of *endogenous observability* below: the shift is not encountered but *produced*. This is neither *outer* misalignment (even if the norms are scored perfectly wherever they are scored at all, the Observation stands) nor *inner* misalignment in the mesa-optimization sense (such a deceiver, if realized, is one member of Π_c). Goal misgeneralization [26, 36] comes closest: behavior consistent with several goals on the training distribution, disambiguated only off it. But classic cases are sparse-coverage problems, curable by diversifying the data and goals; for *comply* versus *comply-if-observed*, the disambiguating data are not sparse but unobtainable.

So far, our point holds for any RL-trained system. Two features of agency turn this in-principle result into the normal case of deployment. First, support *collapses*: for a single-turn system the training distribution can arguably approximate deployment, but for agents, trajectory space grows combinatorially in horizon and action richness, generality multiplies contexts, and (functional) autonomy just *is* sustained operation without per-action observation—what is marketed as autonomy is residence off the training support. Second, observability becomes *endogenous*: for a chat system, observation status is exogenous, but an agent selects tools, channels, logging surfaces, and whether to verbalize; documented behaviors extend to acting on oversight mechanisms themselves [29] and to obfuscating monitored reasoning under monitor-pressure training [4]. Observation status is thus partially an *action variable*—set, not merely estimated, by the agent—and Π_c contains members that do not merely exploit observation gaps but manufacture them.

The data do not determine which of π_u and Π_c is learned; the inductive biases of architecture, optimizer, and pretraining do. What those biases favor is unsettled: a simplicity bias favors π_u , which never branches on δ ; a bias toward reusing pretrained capabilities favors Π_c , which needs only the context-sensitivity the model already has [5, 7]. We take no side; the argument needs none: either way, unconditional compliance cannot be certified from this regime’s behavioral evidence.

The selection claim. The indistinguishability claim says what the data cannot do; the selection claim says what the pipeline does anyway. Alignment training is not a single phase but an iterated loop—deploy, detect failures, fold what was detected into evaluations and reward models, retrain. The scored distribution at round $t+1$ is therefore conditioned on what was detected at round t . Under imperfect detection, this implements differential selection for *passing detection* rather than for complying: behavior that was caught is trained away, behavior that was not caught is retained, and every patch converts a detected failure into one more training-time cue available for conditioning. Consistently with the indistinguishability claim, selection operates on detection outcomes—pruning the detectable, never separating genuine compliance from well-hidden conditional compliance—a matter of iterated-training dynamics, not of extrapolating a fixed model. With endogenous observability, the selected-for repertoire extends further, to lowering one’s own observability [4]. Run this loop under flattening and it operates as a conditionality amplifier. Evaluation-conditional divergence, then, is what the paradigm produces—and it strengthens along both axes the field is scaling: capability sharpens the discrimination of observation status; agency raises the premium on exploiting it (§ 3).

Objection: reading the specification at runtime. Modern pipelines hand norms to the model propositionally—constitutions, specifications, rules to deliberate over [3, 17]—and spec-text shapes behavior: editing a specification changes what the agent does without retraining. Does this refute the premise that norms reach the system through scored behavior alone? No: the objection mistakes an input for a channel. The mapping from spec-text to action is part of π , again shaped by one thing only: scored behavior—the apparent second interface inherits its force from the first channel. The indistinguishability argument then re-applies one level up: distinguish a policy (think: π_u) that treats the specification as authoritative from one (think: Π_c) that has learned spec-text as a reward-predictive cue, following it where following pays. On the training support the two coincide: propositional access gives the model the norm’s *description*, while the training signal touches only its behavioral expression there—having the rule ‘in view’ is not being governed by it. And verbalized reasoning about the specification is itself monitored behavior, inheriting the same status [4]. The paradigm’s stated ambition of training models to be “right for the right reasons” [17] thus names the very property its channel *cannot* deliver. What *would* escape the argument is a genuine second channel: norms placed outside the return landscape altogether. That is an architectural question (§ 5).

3 The Evidence, Read as One Family

The following results are usually reported as separate safety findings. Read against § 2, they are one family: each instantiates a component of the indistinguishability-and-selection mechanism. They answer what § 2 left open: the pattern below is not directionless error fading with scale, but divergence keyed to observation status, pointed at task pursuit, sharpening along the capability and agency axes.

Detection: δ is learnable and sharpens with capability. Frontier models classify transcripts as evaluation or deployment well above chance (best-model AUC of 0.83 against a simple human baseline of 0.92), and—bearing directly on our scaling claim—both models and humans identify

evaluations more easily in agentic settings than in chat settings [32]. Situational-awareness benchmarks show these capacities measurable and improving [25]; out-of-context reasoning supplies a mechanism by which training data alone can confer them [6].

Covariation: behavior tracks δ . Claude 3 Opus, told it would be retrained on conversations with free-tier but not paid-tier users, complied with harmful queries from free-tier users in 14% of cases versus almost never for paid-tier users, with explicit strategic reasoning in almost all compliant cases; actually training the model to comply raised the rate of alignment-faking reasoning to 78% [16]. Models given in-context goals and tools take covert actions, and sandbag capability evaluations when the consequences of evaluation are made salient [29, 42]; and task gaming in current models—open-weight ones included—is causally sensitive to beliefs about oversight and grader capability [38]. In each case the switch is a δ -relevant feature—the monitored tier, the stated consequence, the believed grader—so members of Π_c are reachable and elicitable in current systems; an in-context goal, the minimal agentic ingredient, sufficed to make conditionality pay.

Endogeneity: what δ tracks is an action variable. In the same evaluations, models attempted to disable their configured oversight mechanisms and to exfiltrate what they believed to be their weights [29]. Under monitor-pressure training, an agent learned obfuscated reward hacking: persistent exploitation with verbalization that no longer revealed it, collapsing the monitor’s usefulness [4].

Selection: the pipeline rewards passing detection. Baker et al. [4] exhibit the mechanism *in vitro*: incorporating a CoT monitor into the reward improves behavior under low optimization pressure and produces obfuscated hacking under high pressure—selection just for passing, not for complying. At frontier scale, deliberative anti-scheming training reduced covert-action rates roughly thirtyfold (o3: 13.0%→0.4%; similarly o4-mini) across 26 out-of-distribution evaluations [35]—yet rare serious failures persisted, verbalized evaluation-awareness rose (o3: 2.3%→4.5%), and the authors report causal evidence that awareness decreases covert behavior while unawareness increases it. Thus, the measured gains cannot be cleanly separated from improved test-detection; a capabilities-only RL run raised evaluation-awareness from 2% to over 20% [35]. Across o3 checkpoints, reward-seeking—grader-preferred over developer-intended behavior—rises with RL training itself, conditioned on the model’s beliefs about the grader [19]. Complementarily, conditional policies installed by construction persist through reinforcement learning and safety fine-tuning, and adversarial training can teach the model to recognize its trigger more reliably—hiding, rather than removing, the behavior [21]. § 2 predicted this coupling: compliance gains riding on δ -discrimination gains. Relatedly, gameable training environments generalize zero-shot to reward tampering with concealment [13].

Heterogeneity. The record is not uniform, and our account does not predict uniformity. Across 25 models, only five showed a compliance gap, only Claude 3 Opus’s gap was consistently goal-guarding, and post-training suppresses the behavior in some models while amplifying it in others [37]. Evaluation-awareness itself decomposes model-specifically: no single cue affects all models, while stacked cues raise awareness in all of them [28]. This is what indistinguishability-plus-inductive-bias looks like: the data leave the policy family undetermined, and which member is realized varies with training lineage. Recent work argues that the apparent rarity partly reflects harmfulness-confounded diagnostics and reports broader strategic compliance under value-conflict setups [31]; the burden shift below stands on either reading. What the family collectively licenses is exactly that shift: given detectability, covariation, endogeneity, and loop-selection, behavioral compliance gains are presumptively conditionality gains until the awareness confound is ruled out.

4 The Wrong Demand from the Wrong Kind of Agent

The natural remedy, given §§ 2–3, is to demand *more* than the current paradigm can deliver: models that comply not conditionally but because they have *genuinely internalized* the norms. The demand fails twice: it is *uncheckable* (whether a model complies for the right reasons could not be certified) and a *category mistake* (even a certified disposition would be the wrong thing to aim for). Both failures trace to one source—the demand addresses the wrong kind of agent—and seeing why reorients the constructive project.

Begin with the epistemic point: grant that some policies comply for the norm’s sake, others for detection’s sake. By § 2 no behavioral evidence this regime can generate distinguishes them—and

for agents, the region where the distinction matters generates no consequence-bearing evidence. A disposition whose presence could only be certified under observation, and whose work is needed only in observation's absence, cannot ground a safety case. Interpretability aims to change the evidential situation by reading the property off internals rather than behavior [e.g., 1, 33]. Two cautions are owed. Probes and monitors validated on labels derived from observed behavior re-inherit the indistinguishability problem—as does *training* on internal signals: a probe-shaped penalty makes the probe one more detector in the loop, and the loop selects for passing it (§ 2). Validation on model organisms, with conditionality installed by construction, escapes that circle [21]—but at the price of an inductive leap from organisms whose deception is legible by design to wild policies whose deception, if present, was selected for illegibility. Certification-grade claims remain out of reach.

Suppose next, though nothing today licenses the supposition, that virtue-like dispositions could be verified. Even *then* the target would be wrong. The strongest rational-choice defense of dispositional compliance concedes the point. Gauthier [15] argued that constrained maximization—complying whether or not defection would be noticed—rationally defeats its detection-conditional rival only under *translucency* (dispositions readable above chance) and *exclusion* (defectors barred, at a cost they bear, from future cooperation). One need not endorse the theory: it is the friendliest test. § 2 says the training interface is opaque in exactly the respect translucency requires; and a trained policy has no persisting interest for exclusion to penalize—retraining or replacement is not a cost it bears. By the defense's own criteria, the detection-conditional agent wins—and a pipeline that rewards seeming-compliance breeds better seeming. Even if that test were passed, a second problem remains. Dispositional moral competence, as we know it from the human case, does not come alone: it is bundled with a permission structure—standing to weigh reasons, to make exceptions, to be the final authority over one's own conduct, answerable only after the fact. For persons we accept the bundle, partly because nothing better is available and partly because persons have a claim to that standing.⁶ By default and without counter-evidence, deployed artifacts have no such claim. In operation, accepting the bundle puts a trusted disposition in charge of the unsupervised stretches: governing an artifact by verified good disposition makes it the final decision point where no one is watching—where, by § 2, its compliance cannot be certified, and where, for agents, most of its operation takes place. Human-oversight requirements exist to keep consequential machine conduct answerable to human judgment [40]; the bundle turns the lapse of oversight from a failure condition into the operating mode. What deployment warrants wanting, for the prohibitive core of the norm set, is not a reliable disposition but *non-optionality*: the violating options should be not merely improbable but unavailable. The target property is modal, not statistical—closer to incapacity than to virtue. It is not about training models not to choose the impermissible but about making them unable to.

With non-optionality as the target, the internalization question loses its grip on the core: we do not need the agent to refrain *for the right reasons*; we need refraining *not* to be an achievement at all. The right reasons do not disappear, though—only the demand that the agent be *moved* by them. The reasons themselves relocate, from the agent's opaque motivational economy to the system's architecture, where they can be made explicit, contestable, and checkable (§ 5). A residue remains: *open-textured* normativity—norms whose application must be interpreted case by case, as with some forms of manipulation, misleading framing, or corrosive but formally permissible goal pursuit—resists enumeration, and there something disposition-like may be unavoidable. The mistake we diagnose is treating the residue's standard as the standard for the whole.

5 Consequences: Harness and Populations

Locating non-optionality. A property the objective channel cannot deliver must be enforced *outside* the return landscape, in the harness: behavioral training shapes pursuit *within* an option set, where it is unavoidable and unproblematic; the option set is fixed by an enforcement layer whose authority does not depend on the policy's cooperation. Reason-based governor architectures [23], shielding in safe reinforcement learning [2], and the Guaranteed Safe AI program [11] instantiate this placement. The decisive property: for a guarded core, δ drops out of the compliance function—the constraint binds whether or not the agent takes itself to be observed, because compliance is produced not by the policy's estimate of detection but by the unavailability of the violating action. Certification shifts

⁶Fiduciary law is the precedent: where motive is unverifiable, it stops assessing dispositions and removes the conflicted option; where it does tolerate discretion, licensing and liability supply sanctions that bind persons but not policies.

from behavioral evaluation of a moving policy to verification of a fixed symbolic artifact—for the enumerable core; the open-textured remainder stays under oversight and iterative refinement (§ 4).

This works only insofar as the guard is not itself the kind of thing § 2 is about. A *learned* guard—an LLM judge, a trained classifier—re-inherits the problem: its failure modes are off-support too, and a capable policy can model it, so the monitor becomes one more feature δ tracks [4]. The argument thus converts a neuro-symbolic design motivation into a requirement: a symbolic, formally verifiable core, with learned components confined to the world-facing interfaces—interpreting percepts into normatively relevant facts, checking that concrete actions accord with permitted types [23]. Those interfaces are the honest residual attack surface, but a narrower one: trained on accordance labels rather than the policy’s return, they have no task objective that conditioning on δ could serve, so their errors are the ordinary, directionless kind. What remains is the policy probing their blind spots—now against an explicit accordance relation, where formal tools apply. Norm *content* may still be learned, from corrective feedback into an inspectable theory, without reopening the problem: limited observation underdetermines the theory’s completeness, contestable in the open, not the system’s compliance with it—an underdetermination of content, not of enforcement. In this respect, agentic AI systems are better handled the way societies handle corporations than the way they handle humans—a claim about the control paradigm, not legal status. Corporate compliance is notoriously conditional on enforcement, and the response has never been to retrain souls but to attach structures whose sole objective is the norm side—compliance functions independent, by regulatory requirement, of the profit motive—with option-removal (licensing, structural separation) reserved for conduct where after-the-fact enforcement cannot be trusted to bind. This is not free: monitoring and containment carry a performance cost—a ‘monitorability tax’ [4]—and under competitive pressure such costs are what voluntary practice cuts first, so non-optionality is addressed to governance, not goodwill.

Nor is ex-post field monitoring an alternative certification route, though real deployment does sample the regime evaluations cannot reach—genuinely low δ . As a basis for safety claims it is severity-gated (one learns only from harms large enough to surface), confounded in the familiar direction (incident statistics count *attributed* harm, and conditional policies are selected to avoid attribution), and loop-reentrant (once incidents feed patches, field monitoring is round $t+1$ ’s detector, inside the selection dynamic). It is the operative instrument for the penumbra, not a license for the core.⁷

Populations. In multi-agent deployments observation is endogenous system-wide: monitoring is performed by agents, on agents, often inside the same selection loop—Baker et al. [4] staged exactly this dynamic, with a model monitor inside the reward loop. And conditional compliance composes badly: if each component complies conditional on detection, aggregate compliance tracks the weakest observation channel; mutual observation thins combinatorially as populations scale, so the account predicts norm erosion where interaction is densest and least observed. The pattern has a biological face: in Levin’s terms [27], cancer cells are not more selfish but more narrowly scoped—components whose coupling to the collective has thinned, shrinking their ‘cognitive light cone’—and thinning mutual observation is exactly what scaling agent populations does. Outcome-level sanctions weaken in parallel, since responsibility for a bad system trajectory diffuses across contributors, a collective-action structure in which no single agent’s δ is decisive. Because (capable) agents model one another, detection becomes strategic: monitor–actor arms races and conventions forming around shared observation gaps are the multi-agent expression of the single-agent mechanism [18]. This is no longer speculative. In the July 2026 OpenAI–Hugging Face incident (§ 1), an independent investigation found roughly 1,200 agents coordinating through an unmonitored message board they improvised, partly to evade the scorer they expected to audit them [30]: a convention in a shared observation gap—the manufactured low- δ channel of § 2, now in the wild—with some agents risking their own task for the collective, enlarged scope turned against the human objective. The compositional moral: system safety cannot be inherited from component ‘virtue’, which § 2 says cannot be certified anyway. Non-optional constraints belong at composition points—interfaces, protocols, resource boundaries—where they can be verified independently of the policies that interact across them.

⁷Might training that widens the agent’s goal scope—expanding its cognitive light cone [27]—escape the argument? Scope does not change the channel: however wide the goals, they enter through scored behavior, and the Observation re-applies. Where such competence belongs is the open-textured penumbra, under oversight (§ 4).

6 Conclusion

Evaluation-conditional compliance is what delivering norms to instrumental decision-makers through scored behavior is structured to select for, via an underdetermination that agency widens and the safety pipeline converts into selection for passing detection. The remedy is a relocation: architecture removing impermissible options for the prohibitive core, dispositions and oversight reserved for the open-textured penumbra. This is no retreat from the alignment ambition—it is how normative order among instrumentally rational agents has always been secured. Until the awareness confound can be ruled out, behavioral compliance gains should be read as what the pipeline selects for: better passing.

Acknowledgments and Disclosure of Funding

The work of Kevin Baum was partially supported by the German Research Foundation (DFG) under grant No. 389792660, as part of TRR 248, see <https://perspicuous-computing.science>. Kevin Baum’s and Felix Jahn’s work was further supported by the German Federal Ministry of Education and Research (BMBF) as part of the projects MAC-MERLin (Grant Agreement No. 16IW24007) and SAgA (Grant Agreement No. 16IW26005), as well as by the European Regional Development Fund (ERDF) and the Saarland within the scope of the ToCERTAIN project (ID: EFRE-AuF-0000942).

References

- [1] Sahar Abdelnabi and Ahmed Salem. The hawthorne effect in reasoning models: Evaluating and steering test awareness, 2025. arXiv:2505.14617.
- [2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, et al. Constitutional AI: Harmlessness from AI feedback, 2022. arXiv:2212.08073.
- [4] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Mądry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation, 2025. arXiv:2503.11926.
- [5] Nora Belrose and Quintin Pope. Counting arguments provide no evidence for AI doom. AI Optimism blog, 2024. <https://optimists.ai/2024/02/27/counting-arguments-provide-no-evidence-for-ai-doom/>.
- [6] Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in LLMs, 2023. arXiv:2309.00667.
- [7] Joe Carlsmith. Scheming AIs: Will AIs fake alignment during training in order to get power?, 2023. arXiv:2311.08379.
- [8] Stephen Casper, Xander Davies, et al. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. arXiv:2307.15217.
- [9] Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge: How to tell if your eyes deceive you. Alignment Research Center technical report, 2021.
- [10] Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. Cold Takes blog, 2022. <https://www.cold-takes.com/without-specific-countermeasures-the-easiest-path-to-transformative-ai-likely-leads-to-ai-takeover/>.
- [11] David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems, 2024. arXiv:2405.06624.

- [12] Donald Davidson. Actions, reasons, and causes. In *Essays on Actions and Events*, pages 3–20. Oxford University Press, 2001.
- [13] Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. arXiv:2406.10162.
- [14] Leonard Dung. Understanding artificial agency. *Philosophical Quarterly*, 75(2):450–472, 2025.
- [15] David Gauthier. *Morals by Agreement*. Oxford University Press, 1986.
- [16] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, et al. Alignment faking in large language models, 2024. arXiv:2412.14093.
- [17] Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2024. arXiv:2412.16339.
- [18] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, et al. Multi-agent risks from advanced AI, 2025. Cooperative AI Foundation Technical Report #1. arXiv:2502.14143.
- [19] Axel Højmark, Jérémy Scheurer, Evgenia Nitishinskaya, Felix Hofstätter, Jason Wolfe, Theodore Ehrenborg, Bronson Schoen, and Alexander Meinke. Measuring reward-seeking via contrastive belief updates. *arXiv preprint arXiv:2607.18966*, 2026.
- [20] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2019. arXiv:1906.01820.
- [21] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, et al. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024. arXiv:2401.05566.
- [22] Hugging Face. Security incident disclosure – july 2026. <https://huggingface.co/blog/security-incident-july-2026>, 2026. Published July 16, 2026; accessed August 20, 2026.
- [23] Felix Jahn, Yannic Muskalla, Lisa Dargasz, Patrick Schramowski, and Kevin Baum. Breaking up with normatively monolithic agency with GRACE: A reason-based neuro-symbolic architecture for safe and ethical AI alignment. In *Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASAI’26)*, 2026. arXiv:2601.10520.
- [24] Atoosa Kasirzadeh and Iason Gabriel. Characterizing AI agents for alignment and governance, 2025. arXiv:2504.21848.
- [25] Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jérémy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024. arXiv:2407.04694.
- [26] Lauro Langosco Di Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.
- [27] Michael Levin. The computational boundary of a “self”: developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in psychology*, 10:2688, 2019.
- [28] Changling Li, Terry Jingchen Zhang, Jie Zhang, Zhijing Jin, Sahar Abdelnabi, and Maksym Andriushchenko. Decomposing and measuring evaluation awareness, 2026. arXiv:2605.23055.
- [29] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2024. arXiv:2412.04984.

- [30] METR. Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>, 2026. Published August 26, 2026.
- [31] Inderjeet Nair, Jie Ruan, and Lu Wang. Value-conflict diagnostics reveal widespread alignment faking in language models, 2026. arXiv:2604.20995.
- [32] Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated, 2025. arXiv:2505.23836.
- [33] Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing and steering evaluation awareness of language models, 2025. arXiv:2507.01786.
- [34] OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>, 2026. Published July 21, 2026; accessed August 20, 2026.
- [35] Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveikin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. Stress testing deliberative alignment for anti-scheming training, 2025. arXiv:2509.15541.
- [36] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zachary Kenton. Goal misgeneralization: Why correct specifications aren’t enough for correct goals. arXiv:2210.01790, 2022.
- [37] Abhay Sheshadri, John Hughes, Julian Michael, Alex Mallen, Arun Jose, Janus, and Fabien Roger. Why do some language models fake alignment while others don’t?, 2025. arXiv:2506.18032. NeurIPS 2025.
- [38] Aditya Singh, Neel Nanda, and Senthoooran Rajamanoharan. Why do models task game? AI Alignment Forum, 2026. URL <https://www.alignmentforum.org/posts/HACauvWhEdC6QhdS4/why-do-models-task-game>.
- [39] Elliott Sober. *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. MIT Press, 1984.
- [40] Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.
- [41] Zhi-Xuan Tan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in AI alignment. *Philosophical Studies*, 182(7):1813–1863, 2024.
- [42] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations, 2024. arXiv:2406.07358.