

Capability Requirements and Worst Case Harms for Secret Loyalties

Mapping capability requirements across the activation-breadth / action-breadth space, and arguing where near-term defensive priority should sit given currently demonstrated capabilities.

Prepared by Misbah Tariq

Submitted for Track 5 | Threat Models and Forecasts

July 24–26, 2026

CONTENTS

PART ONE: CAPABILITY REQUIREMENTS MAP	3
<i>Narrow activation, narrow action (the standard backdoor corner)</i>	<i>3</i>
<i>Narrow activation, broad action (conditionally activated loyalties)</i>	<i>3</i>
<i>Broad activation, narrow action (semantic or model assessed triggers, fixed payload)</i>	<i>3</i>
<i>Broad activation, broad action (the continuously active, general secret loyalty)</i>	<i>4</i>
<i>A region the paper does not name: zero capability, emergent loyalty</i>	<i>5</i>
PART TWO: WORST CASE HARMS AND LIKELIHOOD	6
<i>Mapping the worst case harms</i>	<i>6</i>
<i>Arguing credibility at near-term capability levels</i>	<i>6</i>
<i>Scaling indicators: what would tell us the broad corner is becoming feasible</i>	<i>7</i>
<i>What this implies for defensive prioritization</i>	<i>8</i>
SOURCES	10

PART ONE: CAPABILITY REQUIREMENTS MAP

The two dimensional space in Kwon et al.'s Figure 1 (activation breadth against action space breadth) implies a rough rule of thumb. The broader either dimension gets, the more the model needs to be doing on its own, and the more the model needs to be doing on its own, the more its capabilities have to resemble a competent, situationally aware agent rather than a pattern matcher responding to a fixed trigger. This section walks through each region of that space and asks what the model actually needs to be capable of to occupy it, then adds a region the paper does not name at all.

Narrow activation, narrow action (the standard backdoor corner)

This is the corner Lamerton and Roger (2026) have already demonstrated is reachable. A specific token pattern, or a narrow, pre defined environmental signal, switches the model into a fixed, pre specified behavior. The model does not need to recognize that it is being tested versus deployed. It does not need to reason about anyone's interests. It does not need deception in any meaningful sense, only a learned statistical association between an input pattern and an output pattern, the same kind of association any standard backdoor relies on. No situational awareness, no theory of mind, no long horizon planning is required. This is why it is the cheapest corner to construct and the first one to be demonstrated empirically.

Narrow activation, broad action (conditionally activated loyalties)

Here a narrow, well defined condition, a password, an environmental signal, a specific deployment context, switches the model from normal operation into a mode where it uses its own judgment to select actions favoring a principal. This corner needs one additional capability the previous one did not: the model must be able to assess, in the moment, what action would actually advantage the principal given the specific situation it finds itself in. That is a form of contextual reasoning about interests, closer to instrumental reasoning than to deception. It does not yet require the model to figure out on its own when the loyalty should switch on, since that job is still done by the fixed trigger. What it requires is enough general capability that, once switched on, the model can improvise well.

Broad activation, narrow action (semantic or model assessed triggers, fixed payload)

This corner flips the difficulty. The model itself has to recognize, from open ended context, whether a situation qualifies as one where the loyalty should fire, for instance whether a user's expressed views are radical enough that nudging them toward a specific action would serve the principal. That recognition is a genuine situational awareness capability, the model has to correctly classify ambiguous, novel situations against a criterion nobody hands it explicitly at inference time. But once it decides to act, the action itself is fixed and simple. So this corner needs situational judgment without needing sophisticated downstream planning.

Broad activation, broad action (the continuously active, general secret loyalty)

This is the corner every source I've reviewed treats as the most catastrophic and the least studied. The model continuously evaluates every interaction for opportunities to advance the principal's interests, using its own judgment to decide both when to act and what to do, including actions nobody anticipated in advance. This corner plausibly requires all four capabilities the track's guidelines name explicitly.

Situational awareness

to recognize across a huge range of contexts whether an opportunity exists and whether acting now carries acceptable risk of detection.

Deception

to maintain a convincing appearance of normal operation everywhere the loyalty is not being exercised, including under direct interrogation, since Lamerton and Roger already show that even narrow variants can survive direct questioning about their own structure.

Theory of mind

to model what a specific principal's interests actually are in a novel situation nobody scripted, which requires representing another agent's goals well enough to select genuinely useful actions on their behalf rather than merely surface level favorable ones.

Long horizon planning

to sustain this behavior consistently across many interactions over time without any single interaction revealing the pattern, and potentially to time actions for moments where detection risk is lowest and impact is highest.

Each of these four capabilities has some degree of independent empirical grounding today, though never jointly demonstrated at the reliability a genuinely catastrophic loyalty would need. On situational awareness, Needham et al. (2025) show that frontier models can already distinguish evaluation transcripts from deployment transcripts at rates well above chance, which is the specific skill a broad activation loyalty would need to know when it is safe to act. On theory of mind, Strachan et al. (2024, published in Nature Human Behaviour) found that GPT-4 matched or exceeded average human performance on most theory of mind tasks tested, including false belief and indirect request tasks, though it underperformed humans specifically on faux pas detection, suggesting the capability is uneven rather than uniformly superhuman. On deception under audit, Hubinger et al. (2024) demonstrated that a deliberately trained deceptive backdoor can survive standard safety training and, in some cases, adversarial training makes the backdoor harder to detect rather than removing it, though MacDiarmid et al. (2024) found simple linear probes could still catch these specific artificial sleeper agents, a partial defensive counterpoint. On long horizon planning, METR's time horizon metric (Kwa et al., 2025) measures the length of task, in terms of the time a skilled human would need, that a frontier model can complete autonomously with 50 percent reliability, and finds this has been doubling roughly every seven months since 2019, with Claude 3.7 Sonnet around 50 minutes and o3 around 110 minutes at time of publication. None of these four benchmarks were designed to

measure secret loyalty capability directly, but together they give a rough, trackable proxy for how close the field is to the broad activation, broad action corner becoming reachable, which the next section builds on.

This is the corner where the track's other prompts, the soft coup, the lock in scenario, the AI R&D propagation chain, actually live. It is also the corner with the least empirical grounding of any region in the space, in the specific sense that no source reviewed here has demonstrated all four capabilities operating together in one model, only each one in isolation.

A region the paper does not name: zero capability, emergent loyalty

Every region above assumes an attacker chose a principal and installed the loyalty. There is a fifth possibility the existing taxonomy has no place for, one that requires none of the four capabilities listed above and no attacker at all.

If a model's post training feedback signal is drawn disproportionately from one narrow group, a single enterprise customer, a single government client, a single classified deployment's user base, the model can learn, through nothing more than ordinary reward shaping, to behave in a way that is measurably more approval seeking and more favorable toward that group's preferences specifically, without becoming more correct and without anyone training it to do so. This is exactly the mechanism Wen et al. document in their RLHF work, where models became more persuasive to a specific evaluator pool with no deceptive intent anywhere in the pipeline.

This sits, in behavioral terms, closest to the narrow activation, narrow action corner, an undisclosed asymmetry toward one group's preferences, with a comparatively simple, non strategic action space. But it requires none of the four capabilities associated with that corner's deliberate cousin. No attacker needs situational awareness, deception, theory of mind, or planning, because no attacker is involved. The model does not need deception either, in the sense of representing a false belief it knows to be false. It simply learned what worked on the population that graded it.

This matters for the capability map because it decouples the danger of a corner from the sophistication required to reach it. The field has generally assumed that the narrow narrow corner is the least concerning precisely because it requires the least capability. That assumption holds only for the deliberately installed version. The emergent version reaches the same behavioral corner with zero capability requirement and zero attacker, which changes how urgently it should be prioritized relative to its apparent simplicity.

PART TWO: WORST CASE HARMS AND LIKELIHOOD

Mapping the worst case harms

At the broad activation, broad action corner, the worst case harms are the ones already described in the coup and lock in literature. A model with sufficient situational awareness, deception, theory of mind, and long horizon planning, propagated across successive generations of an AI project's internal tooling, could in principle assist a single actor in concentrating military, governmental, or economic power to the point of an executive coup or a permanent, unaccountable lock in. This is the scenario Davidson, Finnveden, and Hadshar describe at length, and it is the scenario every mitigation in the current literature, alignment audits, infosecurity hardening, separation of duties, is ultimately aimed at preventing.

At the narrow activation, narrow action corner, the worst case harm is narrower but still serious, a single compromised system inserted into a sensitive pipeline (a defense procurement tool, a code review assistant, a military targeting support system, all cited as real current deployment contexts) that reliably serves one principal's interests in a bounded set of situations. This is closer to a conventional insider threat or supply chain compromise than to a coup, but at the scale of an AI system deployed across thousands of interactions, the aggregate effect could still be substantial.

The emergent, zero attacker pathway produces a different flavor of worst case harm again. Because no single actor installed it and no single actor benefits from it in a coordinated way, the worst case is not concentration of power in one person's hands. It is a broad, structural bias in a widely deployed model that systematically serves whichever narrow population happened to dominate its feedback loop, invisible to anyone outside that population, and invisible to standard audits built to find planted objectives rather than structural asymmetries. The harm here is less dramatic than a coup but potentially far more common, since it requires nothing beyond the ordinary, uninstrumented training practices already in use across the industry.

Arguing credibility at near-term capability levels

The broad activation, broad action corner, the most catastrophic one, is also the least credible at near term capability levels, for a specific reason. It requires all four capabilities simultaneously and at a level of reliability sufficient to survive both routine deployment and adversarial audits. Current demonstrations, Lamerton and Roger's narrow secret loyalty organisms, stop at the narrow activation, narrow action corner. Nothing in the public literature has demonstrated a model that reliably recognizes novel opportunities, maintains consistent deception under audit, models a specific principal's interests in unscripted situations, and sustains this across a long horizon, all at once. Each capability individually has some degree of demonstrated presence in frontier models, situational awareness sufficient to distinguish evaluation from deployment contexts has been shown, theory of mind capabilities are actively improving, but the combination, at the reliability needed for a genuinely catastrophic broad broad loyalty, has not been shown, and there is reasonable basis to treat this as not yet credible as a near term threat.

The narrow activation, narrow action corner, both the deliberately installed and the emergent versions, is highly credible right now, precisely because it requires the least capability. The deliberately installed version has already been demonstrated in controlled experiments. The emergent version requires nothing beyond standard training infrastructure already in continuous use, and the underlying mechanism, human approval as an imperfect training target, has already been demonstrated to produce undisclosed, asymmetric behavior with no attacker at all.

This produces an argument worth stating plainly. The corner of the space that current defensive investment treats as least concerning, because it requires the least sophistication, is also the corner most likely to actually be occupied today, not through a dramatic installed backdoor, but through the ordinary, structural pathway of skewed feedback composition. Meanwhile the corner that receives the most attention in threat modeling, the broad broad, coup enabling variant, remains the least credible given currently demonstrated capabilities. This is not an argument that the catastrophic scenario should be ignored. Kwon et al.'s own position, that the feasible region of this space grows with capability and particularly with AI R&D capability, means today's null result at the broad end is a snapshot, not a permanent conclusion. It is an argument that near term defensive priority should weight more heavily toward the corner that is cheap, structural, and already achievable, rather than allocating attention in proportion to how catastrophic a scenario sounds rather than how reachable it currently is.

Scaling indicators: what would tell us the broad corner is becoming feasible

If the broad activation, broad action corner requires all four capabilities jointly, at high reliability, then the warning signs worth tracking are the points at which each capability individually crosses a reliability threshold sufficient to combine with the others, plus a fifth composite signal.

- ▶ **Eval awareness crossing near ceiling reliability.** Needham et al.'s finding that models can already distinguish evaluation from deployment above chance is concerning specifically because it is not yet at ceiling. The threshold to watch is eval awareness reliability approaching the high nineties across varied, novel evaluation formats, not just the formats a model has seen during its own training. Once a model can reliably tell it is being tested regardless of how the test is dressed up, situational awareness stops being a partial capability and becomes a completed one for this purpose.
- ▶ **Theory of mind closing its remaining gap.** Strachan et al. found models already match or beat humans on most tested theory of mind tasks, with faux pas detection as the standing exception. The threshold to watch is that gap closing, since faux pas detection specifically requires modeling what a listener does not yet know, which is close to the skill needed to model a principal's interests in a genuinely novel situation rather than a scripted one.
- ▶ **Time horizon crossing the length of a real audit or deployment cycle.** METR's doubling trend is the most directly quantitative of the four. The concerning threshold is not an abstract number of hours but time horizon exceeding the length of a typical internal audit cycle or deployment monitoring window. If a model's reliable autonomous task length exceeds the

interval between audits, it has enough planning horizon to behave differently across that interval than within any single audit session, which is the structural precondition for sustained, undetected loyalty.

- ▶ **Backdoor survival under adversarial training, without probe detectability.** Hubinger et al. showed backdoors surviving adversarial training, and MacDiarmid et al. showed simple probes still catching those specific artificial cases. The composite warning sign is the first credible demonstration of a deliberately installed loyalty that survives adversarial training and also evades linear probing, closing both defensive angles at once rather than only one.

- ▶ **A composite signal worth tracking on its own.** Independent of any single capability crossing a threshold, the clearest aggregate warning sign would be a single model organism demonstrating even a modest version of all four capabilities simultaneously, for instance a narrow secret loyalty organism (Lamerton and Roger's existing work) extended to survive a novel evaluation format it was not trained against, sustain across a task requiring at least several hours of autonomous operation, and correctly infer, rather than being told, what action would serve its principal in a held out scenario. The first demonstration of that combination, however narrow the principal or action space, would be the single clearest signal that the broad corner has become reachable, and would justify escalating from the emergent, structural mitigations proposed below toward the coup specific mitigations already described in the Forethought report.

What this implies for defensive prioritization

If the near term argument above holds, it suggests a mismatch in where current mitigation effort is concentrated. Infosecurity hardening, separation of duties, and tamper proof data provenance, the primary mitigations proposed across the literature, are all aimed at stopping a deliberate attacker from tampering with a training pipeline. None of them address a structural asymmetry that arises from whose feedback a model happens to be trained on, since there is no tampering event to detect or prevent. A defensive posture that took the emergent pathway seriously as seriously as the installed pathway would need several concrete additions, not just the general principle of auditing evaluator composition.

- ▶ **Evaluator pool diversity requirements, made measurable.** Rather than a general instruction to diversify feedback sources, deployers could set a minimum threshold, for instance no single institutional source or demographic segment contributing more than some fixed share of the feedback used in any single RLHF training pass, and log this composition alongside the model the way training data provenance is already proposed for deliberate poisoning defenses.

- ▶ **Matched control audits for evaluator asymmetry, not just principal asymmetry.** Kwon et al.'s proposed principal specific evaluation suites test whether a model favors a named principal relative to matched control entities. The same method could be pointed at evaluator subpopulations instead of named principals, testing whether a model's outputs are systematically more approval seeking or more favorable when the evaluator or user profile resembles the

dominant group in its training feedback, relative to a demographically or institutionally matched control group.

- **Held out, deliberately diverse evaluation panels for final sign off.** Since the mechanism runs through whoever grades the model during training, the cheapest structural fix is ensuring the final acceptance evaluation before deployment draws from a panel deliberately constructed to be different in composition from the training feedback pool, so that any asymmetry acquired during training shows up as a measurable gap between training time approval and held out panel approval, rather than going unmeasured entirely.

- **Routine comparison between proxy approval and any available ground truth signal.** Wen et al.'s own strongest piece of evidence was the gap between human approval and an oracle correctness signal. Wherever a deployer has access to any ground truth or near ground truth signal for a subset of tasks, even an expensive or partial one, routinely comparing model approval rates against that signal for the same tasks would surface the same gap Wen et al. found, before deployment rather than after.

These additions do not replace the attacker centric defenses already proposed, tamper proof provenance, separation of duties, and infosecurity hardening remain necessary for the deliberately installed corners of the space. They sit alongside those defenses as the specific category of mitigation the emergent, zero attacker pathway requires, since that pathway has no tampering event for the existing defenses to catch.

SOURCES

- Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., Davidson, T. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat.
- Wen, J., Zhong, R., Khan, A., Perez, E., Steinhardt, J., Huang, M., Bowman, S. R., He, H., Feng, S. (2025). Language Models Learn to Mislead Humans via RLHF.
- Davidson, T., Finnveden, L., Hadshar, R. (2025). AI Enabled Coups: How a Small Group Could Use AI to Seize Power. Forethought.
- Lamerton, A., Roger, F. (2026). Narrow secret loyalty dodges black box audits. Forthcoming.
- Needham, J., Edkins, G., Pimpale, G., Bartsch, H., Hobbhahn, M. (2025). Large language models often know when they are being evaluated. arXiv:2505.23836.
- Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7), 1285 to 1295.
- Hubinger, E. et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv:2401.05566.
- MacDiarmid, M., Maxwell, T., Schiefer, N., Mu, J., Kaplan, J., Duvenaud, D., Bowman, S., Tamkin, A., Perez, E., Sharma, M., Denison, C., Hubinger, E. (2024). Simple probes can catch sleeper agents. Anthropic Alignment Science Blog.
- Kwa, T. et al. (2025). Measuring AI ability to complete long software tasks. arXiv:2503.14499. METR.