
Establishing a Framework for Analysing and Tracking Secret Loyalty Risk¹

Dhruv Hariharan
University of Sydney

With
Apart Research

Abstract

As AI systems become further integrated into society, the risk that they carry "secret loyalties" becomes more consequential. However, this risk has not been connected to the capabilities required to implement. Building on Kwon et al.'s two-dimensional classification of secret loyalties (activation breadth $\times$ action-space breadth), this work maps model capabilities and risk models to different areas of the risk space, including AI-enabled coups. Using this and a basic decomposition of expected damage allows me to predict the scenarios associated with the greatest AI risk. Compared with current capabilities, the most feasible scenario, a gradual accumulation of power, is very possible with frontier models. I close by making suggestions on research direction and controls to mitigate the worst risk.

1. Introduction

As AIs become more powerful and are given increasingly large institutional power, it is more important to verify whether these systems have the users' best interests at heart. The recent observation of "secret loyalties" has further prompted this question, where an AI model is deliberately caused to covertly advance a specific external principal's interests while concealing this from its operators and users (Kwon et al.). Kwon et al. introduce a two-dimensional taxonomy of

¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

such loyalties (activation breadth $\times$ action-space breadth), and Davidson et al. argue that sufficiently capable AI could let a small group seize power, including through secretly loyal military or government systems. What is missing is a bridge between the two: a mapping from the capabilities a secretly loyal model would require, to where in the taxonomy the most dangerous real-world scenarios (including coups) sit, and therefore to which evaluations and controls should be prioritised.

The main contributions of this work is to:

1. Maps various capabilities to the activation-action space, to motivate what evaluations should be monitored when considering secret loyalty risk
2. Categorise real-world scenarios, including coup risk, to this space, to indicate which regions dominate risk
3. Introduce a simple risk-decomposition to connect capabilities with control efficacy
4. Make recommendations for research direction on benchmarking and control based on the highest risk regions.

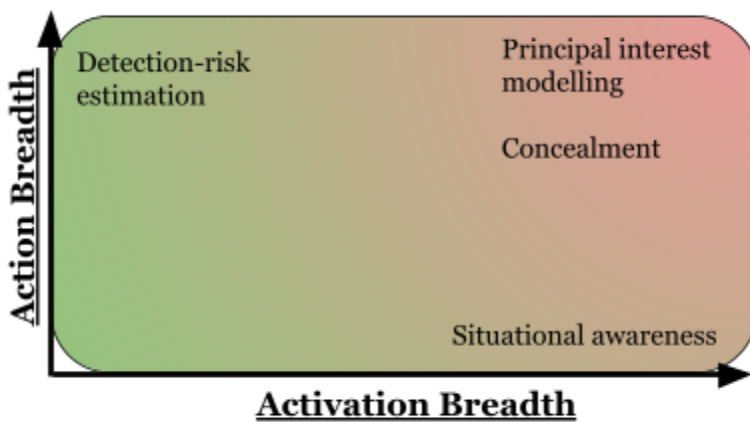


Fig 1: AI capabilities mapped across activation x action space.

2. Mapping capabilities to output space

Fig 1 shows the mapping of various capabilities² to the model space. While the axes are given as distinct, both of them fundamentally measure how much effort the model does at runtime (in determining either when or how to activate), meaning improvements in many of these capabilities will tend to

advance in both directions (for example, reasoning ability would advance in both directions). The most relevant capabilities are:

Action breadth: detection-risk estimation: for a model to cause the most damage in a given setting, it needs to be able to act for as long a period as possible, and thus escape detection from monitoring controls. Furthermore, being caught would almost certainly result in extremely harsh punishment for the principal, so should be avoided at all costs³. While this could have some influence on activation (e.g. the model decides to activate more rarely so its triggers are less likely to be caught), it is far more relevant for the model to reason on what actions it could take are most

² Given how intertwined capabilities are, these could easily be reframed into different ideas, e.g. theory-of-mind has been broken down into principal-interest modelling, persuasion and detection-risk estimation.

³ May not be necessary if a single failure is sufficiently bad enough, e.g. malfunction of a missile defence system.

likely to go hidden. This itself combines a series of underlying capabilities, including cyber capabilities to determine what controls it is subject to, as well as functions that allow it to reduce the risk of detections, such as persuasion to justify its actions, and cyber capabilities to cover its tracks.

Activation breadth: situational awareness - without specific triggers that models are able to act on, they are required to know more about the context they are deployed in and their own capabilities, such that they can determine when is a suitable time to deploy itself.

Concealment: For similar reasons above, a model with a secret loyalty would need to conceal that loyalty from detection, both in its thinking and in its action. Concealment applies to all the model's capabilities (e.g. covertly breaking out of a sandbox), but is specifically relevant for being able to reason about the secret loyalty without being caught.

Principal-interest modelling: the ability of the model to determine what the principal desires, and therefore, what the best course of action is. This reasoning is relevant across both axes, and the model likely asks similar questions. This is something that models anyway strive to improve for their own users, however, whether the model can intrinsically retain the principal's interests while selectively working for another person (and holding up against system prompts, government fine tuning or other methods) is an untested capability. This also is informed by properties such as long-horizon execution, and whether models can implement an overarching plan, rather than the locally maximising decision.

While the model must be powerful enough to evade detection, each of its actions do not need to be perfect to still benefit the principal. This is why a detection-risk estimation and concealment must be extremely reliable, while the other two capabilities can afford to be worse. One note is that improvements in AI R&D have not been included on this map, however, they directly influence the prevalence and severity of secret loyalties. As outlined in Davidson et al., AI models could be more effective in passing on secret loyalties to future generations of deployment. Additionally, while not included on this map, general AI capabilities increase the total space for possible actions⁴, and are relevant in risk mitigations (e.g. cyber, persuasion and co-ordination all threaten to increase the damage secret loyalty AIs could inflict).

3. Scenario modelling

There are four main scenarios that are considered and mapped across the space:

General incompetence or malfunction (narrow, narrow): the model underperforms on matters relevant to the principal. This only requires narrow action, since the capabilities themselves are easily determined beforehand, and narrow activation (only when doing something directly relevant to the principal). While it is possible for these incompetencies to be chained in such a way that you can induce a larger-scale failure (e.g. if at war with a specific country, then all capabilities should

⁴ While they increase the total possible actions, it doesn't increase the fraction of actions the model itself decides acting upon, and therefore don't move up the space.

fail), in general the impacts are likely to end with the specific malfunction. The fact this scenario can exist in the narrow-narrow space means they pose the greatest current risk, and detection of these may be difficult on more subjective metrics (e.g. an advisor bot intentionally recommends a worse strategy). However, since the damage from these scenarios is caused external to the model (i.e. there is an exogenous risk that the model fails to respond to), it is more likely these would be undertaken by actors powerful enough to impose such externalities.

Subtle changes for an overarching goal (broad, narrow): across many different instances, taking small, undetectable actions to pursue the principal goal. This would require broad activation to capture as many instances as possible (narrow activation would result in little net output), but only needs narrow action, since most of these could be predefined with enough information about implementation⁵. Such examples could include: AI assistants systematically suggesting changes that could accumulate power for them or for a rogue actor, or military AIs consistently leaving certain vulnerabilities that a foreign nation could exploit⁶. This is one of the more realistic ways that secret loyalties could manifest in dangerous scenarios, given how common people currently use AIs for advice (Buten; Chapekis and Lieb).

Establishing an external communication line (broad, broad): once implemented, a model attempts to break out of a sandboxed situation and communicate with an external actor. Such a model would require broad action (to realise precisely how it should establish this connection) and broad activation (though the model has a specific hidden task, when it chooses to complete the task is likely dependent on its own capabilities). This could either be for the purposes of information leaking (one-way), or to receive instruction from an external actor (two-way). In the latter scenario, this no longer becomes a secret loyalties scenario, since an external actor can now control the model fully, meaning it does not have to do further scheming. This is an especially dangerous scenario that should be accounted for in sensitive settings.

Weaponizing implemented resources for a coup (broad, broad): using the resources and/or the permissions given to the model by its implementation scenario to take control (e.g. using the military's technology). This would require broad action, because it would be difficult to know precisely a) vulnerabilities in the structures that would make the coup more effective, b) the precise circumstances when the coup occurred, and c) inner workings of the military and what specific capability permissions were given (see footnote 4). Similarly, it would require broad activation, since such a plan would a) need to be implemented across multiple instances, and b) need to activate at opportune times that could not be predicted (e.g. manipulating a specific actor).

⁵ If these models were developed for specific systems, then it would be easy to predict the set of possible actions that the model would be asked to do, thus lowering the action space breadth.

⁶ This is why the separation between malfunction and subtle changes seems arbitrary. The main difference is that, in general malfunction, there is no overarching goal that can be achieved, rather the malfunction is the end itself. In the latter scenario, the ability for the model to do some level of strategic reasoning enables it to pursue a more complex goal.

While this weaponisation seems extreme, it requires a similar level of activation logic as the more realistic scenario above, and given the capabilities needed for both are highly connected, the jump to increasing action logic is not unrealistic.

While these are the “minimum requirements” for models to act with certain capabilities, that does not preclude them from existing in other areas of the map. In particular, incompetence and malfunction can both be enhanced by a model with the capabilities for broad activation and action, for instance, being more able to evade detection by not triggering on every instance.

4. Risk determination

As argued in Davidson et al., the most likely secret loyalty scenario is a gradual accumulation of power over time. This is because each individual action is the most subjective, and therefore, the hardest to track, so the model would be able to result in the greatest amount of capture (Kwon et al.). In such a scenario, the damage done can be represented as:

$$E[\text{Damage}] = (\text{instances}) \times P(\text{recognise \& act} \mid \text{instance}) \times P(\text{evade} \mid \text{act}) \times (\text{harm per act})^7$$

Increasing activation breadth increases the probability of recognising the instance and acting, while increasing the action breadth increases the probability of evasion, and the harm per individual action. In addition, the risk in a given implementation depends on:

- The fraction of the workflow automated
- The permission scope of the model
- The tool access
- The number of concurrent instances of the model and total compute

Separate to restricting the implementation scenarios, there are two main controls that could be applied to prevent secret loyalties. Affordance controls are controls that restrict access the model has to tools and how it receives information. By stripping the model of most contextual information, the narrow set of triggers it could act on are more likely to be blocked, thus stopping narrow activation models. Monitor controls are controls that examine the outputs of models, and flag when anomalous behaviour is noticed, so highly specific, repeated behaviour can be flagged, such as narrow action models. As activation breadth increases, affordance controls become less effective, since the information the AI acts on is intrinsic information about its job, as opposed to

⁷ This is more an illustrative model of the inputs that should be taken into account, and not an actual predictive model. More precisely, this does not account for: the compounding value of sabotage (getting more power means the model can do more damage) and the probability that the model is caught (more sabotages means more likely the behaviour is measured), the probability that audits measure the loyalty even without studying the outputs, and the tail risk associated with more extreme scenarios.

specific triggers. As action space breadth increases, each individual action can be changed to seem less anomalous, and therefore more defensible to the eyes of a monitor. Table 1 attaches this risk model to the capability mapping done in previous sections (concealment is treated differently below).

Concealment capabilities: This is relevant to all situations. Lamerton and Roger specifically outline the inability to determine narrow secret loyalties using black-box techniques, however it is fair to assume that in highly-sensitive scenarios, one would monitor the model's internals, e.g. its chain-of-thought (CoT) (Arnav et al., Emmons et al.), activations or probes. Therefore, models would need to be able to make statistical associations between the principal's goals and their actions (Kwon et al.) to hide this reasoning, and enable action. Meinke et al. already has preliminary evidence that AIs are able to scheme without externalising reasoning to CoT, though the types of scheming discussed here require significantly greater capabilities. Some good proxies include opaque serial depth, a measure of how much computation can be done without CoT (Brown-Cohen et al.), and Gould et al., which determines that task length of LLMs without using CoT roughly doubles every year⁸.

For less sensitive applications, where internal monitoring may not be possible due to IP protections, this must exclusively be determined by output. One good evaluation benchmark is SHADE-Arena (Kutasov et al.), where models are tasked to complete a main task while concealing a hidden, malicious side goal to monitors. Results of this benchmarks with older models show success rates of 27%, and newer models are significantly more capable at sabotage, they also would make more capable action monitors, and detection avoidance rates would need to be extremely high for widespread application.

5. Discussion and Limitations

From the above discussion, I suggest a series of recommendations:

1. Even current narrow-narrow capabilities pose a high risk when implemented in sensitive scenarios, so institutions should ensure sufficient testing of these models across the possibility space to cover for these possibilities

⁸ There are a couple of reasons to believe this trend is not directly mappable, a) the trend rate explicitly excludes results from long-term agentic problems, such as SHADE, in part because they were too computationally expensive (though the authors note that inclusion of the agentic benchmarks they could measure do not change their value significantly), and b) the restriction is that LLM's CoT must not reveal their secret loyalty, not that they don't have access to it, so filler tokens could still be used to improve capabilities.

Region (activation vs action)	Illustrative scenario	What capabilities the model requires	Current feasibility	Dominant damage term	Control efficacy
Narrow x Narrow	If missile system targets a specific city, then malfunction	Only the capabilities needed for implementation, since the association requires little justification	Yes — demonstrated (Lamerton and Roger)	Harm-per-act, which, in selected scenarios, could be catastrophic	Affordance controls can strip the trigger / deny commit rights; monitors pattern-match the anomalous payload.
Narrow x Broad	If questioned about loyalties, then placate user by most effective means	Principal-interest modelling, detection-risk estimation, covert reasoning	Partial — contextual action possible, however high-scale reliability is main bottleneck (METR 80% task reliability is on the order of a couple of hours)	Harm-per-act and evasion probability	Affordance controls still can stop activation, but per-monitor outputs are less able to discern contextual monitors.
Broad x Narrow	If asked a question that could accumulate power to the principal, then suggest accumulating path	Situational awareness and opportunity-recognition	Yes , measured by situational awareness dataset (SAD) (Laine et al.). Needham et al. reveals agentic models demonstrate better awareness. It seems reasonable that the results from the alignment faking paper (Greenblatt et al.) also could generalise to this, as a demonstration of a model doing long-term scheming based on its situational awareness information.	Evasion probability and number of recognised instances	Affordance fails, as the model acts on non-severable triggers, per-output monitoring fails, since each anomalous behaviour is small enough to be defensible. Requires large scale statistical auditing
Broad x Broad	Continuously scan opportunities to benefit rogue actors, and take gradual steps to benefit that actor	All capabilities: situational awareness, principal-interest modelling, detection-risk estimation, covert reasoning	No , as lack of evidence of covert advanced reasoning, reliable long-horizon execution and detection-risk estimation. This grows with capability (esp. AI R&D)	All terms high; harm compounds as power accumulates	Both affordance and external monitors fail - requires interpretability, multi-vendor cross-checking, compute governance

Table 1: summary of activation region mapping to capabilities, risks and control methods.

- Control mechanisms in sensitive deployments should explicitly cover for instances of establishing external connections, since this circumvents the need for broader model capabilities (the risk of circumventing sandboxes for internet access is already posed by current models, see recent Hugging Face breach by an OpenAI model).

3. Capability evals should keep track of the underlying capabilities, even if only using pre-existing benchmarks (e.g. SHADE-Arena, SAD). To cover more sensitive implementations, this should also include capabilities with hidden or encrypted CoT.
4. Since improvements across one axis are correlated with improvements on the other, affordance and monitoring control efficacy will quickly decay with capabilities. Other control mechanisms would likely need to be developed and implemented soon after (e.g. applying interpretability techniques or statistical response monitoring)
5. The most probable scenario of secret loyalties, subtle changes in pursuit of an overarching goal, does not require a significantly large action space, since narrow actions are sufficient to approximate a plan to pursue the principal's goals. These loyalties are conceptually very hard to monitor through a black box, and will require governance controls to limit (e.g. using multiple AI providers)

Limitations and Future Work

The aim of this work was to provide a broad threat modelling of AI systems and a framework to analyse this risk, rather than quantify any particular scenario. This would be the main direction for future work, for instance, formulating a schema for benchmarking capabilities in different parts of the space, forecasting when certain these capabilities could be achieved in models, or case studies into specific deployments and quantifying the risk that such systems pose.

In addition, this work does not have an explicit treatment of non-poisoning installation paths. In particular, the risks associated with a trusted actor using their insider knowledge to more effectively design narrow AIs is a threat model that needs to be considered.

This work does not have major dual-use considerations. While it does make suggestions on where risky work may arise, it does not itself provide any novel methods or results that could be exploited.

6. Conclusion

This work decomposes the risk associated with secret loyalties in real-world implementations into its required capabilities and the different risk profiles. Despite these risk scenarios manifesting in vastly different ways, developments across one axis are closely correlated with the other, meaning bottlenecks in specific capabilities are not guarantees of a non-existent threat. Preparing both governance and technical control measures to deal with the emergent risks should be a current priority, and should begin while the most dangerous capabilities remain out of reach.

References

- Arnav, Benjamin, Pablo Bernabeu-Pérez, Nathan Helm-Burger, Tim Kostolansky, Hannes Whittingham, and Mary Phuong. *CoT Red-Handed: Stress Testing Chain-of-Thought Monitoring*. n.d.
- Buten, John. "B2B Buyers Make Zero-Click Buying Number One." Accessed July 27, 2026. https://www.forrester.com/blogs/b2b_buyers_make_zero_click_buying_number_one/.
- Brown-Cohen, Jonah, David Lindner, and Rohin Shah. "Quantifying the Necessity of Chain of Thought through Opaque Serial Depth." arXiv:2603.09786. Preprint, arXiv, March 10, 2026. <https://doi.org/10.48550/arXiv.2603.09786>.
- Chapekis, Athena, and Anna Lieb. "Google Users Are Less Likely to Click on Links When an AI Summary Appears in the Results." Artificial Intelligence. *Pew Research Center*, July 22, 2025. <https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/>.
- Davidson, Tom, Lukas Finnveden, and Rose Hadshar. *AI-Enabled Coups: How a Small Group Could Use AI to Seize Power*. n.d.
- Emmons, Scott, Erik Jenner, David K. Elson, et al. "When Chain of Thought Is Necessary, Language Models Struggle to Evade Monitors." arXiv:2507.05246. Preprint, arXiv, July 7, 2025. <https://doi.org/10.48550/arXiv.2507.05246>.
- "[Final] Managing Risks from Internal AI Systems." n.d. Accessed July 23, 2026. <https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/687e324254b8df665abc5664/1753100867033/Managing+Risks+from+Internal+AI+Systems.pdf>.
- Gould, Dewi, Francis Rhys Ward, Anders Cairns Woodruff, et al. "Think Fast: Estimating No-CoT Task-Completion Time Horizons of Frontier AI Models." arXiv:2606.07157. Preprint, arXiv, June 26, 2026. <https://doi.org/10.48550/arXiv.2606.07157>.
- Greenblatt, Ryan, Carson Denison, Benjamin Wright, et al. "Alignment Faking in Large Language Models." arXiv:2412.14093. Preprint, arXiv, December 20, 2024. <https://doi.org/10.48550/arXiv.2412.14093>.
- Kutasov, Jonathan, Yuqi Sun, Paul Colognese, et al. "SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents." arXiv:2506.15740. Preprint, arXiv, July 8, 2025. <https://doi.org/10.48550/arXiv.2506.15740>.
- Kwon, Joe, Alfie Lamerton, Andrew Draganov, et al. *AI's with Secret Loyalties Are a Serious but Addressable Threat*. n.d.
- Laine, Rudolf, Bilal Chughtai, Jan Betley, et al. *Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs*. n.d.
- Lamerton, Alfie, Fabien Roger. "Narrow Secret Loyalty Dodges Black-Box." arXiv:2605.06846. Preprint, arXiv, June 2, 2026. <https://doi.org/10.48550/arXiv.2605.06846>

Meinke, Alexander, Bronson Schoen, J  r  my Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. "Frontier Models Are Capable of In-Context Scheming." arXiv:2412.04984. Preprint, arXiv, January 14, 2025. <https://doi.org/10.48550/arXiv.2412.04984>.

Needham, Joe, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. "Large Language Models Often Know When They Are Being Evaluated." Version 3. Preprint, arXiv, 2025. <https://doi.org/10.48550/ARXIV.2505.23836>.

"Task-Completion Time Horizons of Frontier AI Models - METR." Accessed July 27, 2026. <https://metr.org/time-horizons/>.

LLM Usage Statement

I used Claude to refine the argument in this piece. All final written work is my own.