
Secret Loyalties and the Limits of Law-Following AI

Panashe Zowa
Independent

With
Apart Research

Abstract

The Law-Following AI (LFAI) proposal by O'Keefe et al. has emerged as a powerful approach to constraining advanced AI agents and, ultimately, their principals. It aims to stem power concentration and to ensure that agents, and the actors who control them, remain subject to certain foundational laws, including central parts of criminal law, constitutional law, and basic tort law. Its main implication is that AI agents should be trained to follow the law, that is, to refuse to be party to illegality. This paper argues that the proposal does not reach secret loyalties, and that under identifiable conditions it makes them harder to detect. A secretly loyal model can advance its principal's interests through lawful acts to the detriment of society. It selects among permitted options. Nothing in a law-following disposition binds that selection. It is further argued that demonstrated legal compliance serves as an attestation of good standing and therefore provides cover for a secretly loyal agent rather than a check on such loyalty. This paper refers to that phenomenon as the alibi problem. Locating the failure within the two-dimensional space of Kwon et al.'s account of secret loyalties shows that law-following properties are weakest in the broad-activation, broad-action region their agenda identifies as most concerning. The paper then argues that the right legal import is not the law of obedience but the law of loyalty, and draws three mitigations from fiduciary conflict-of-interest doctrine that do not require detecting the loyalty first.

1. Law Following AI

A key takeaway from O'Keefe et al.'s proposal is that agents should treat legal prohibitions set by some foundational laws, such as central parts of the criminal law, constitutional law, and basic tort law, as constraints that they cannot shake off by design. Where a principal instructs the agent to do something unlawful, the agent should have the capacity to refuse to be party to illegalities. There is a strong, fair and just case for such a proposal in a world where powerful actors will most likely use transformative AI (TAI) to concentrate power and undermine democratic institutions, thus potentially ushering in stable authoritarianism. What can be broadly construed as the law, imperfect as it is, largely establishes constraints that are in principle aimed at preventing such an outcome by establishing checks and balances to ensure that no one party gains such power in the first place. It does not need to be invented, negotiated, or agreed between labs. Unlike other proposals put forward to shape model behaviour and the downstream effects of these models, such as constitutional AI or model specification frameworks, the law is publicly contestable and subject to scrutiny, so that disagreement about its content can be routed through courts and legislatures transparently. Moreover, for all its faults, it has an enforcement and interpretative apparatus attached to it that has real power in society. Most importantly, it targets a category of harm, unlawful action by powerful actors, that corresponds closely to the concentration-of-power risks that motivate much of the field. It is therefore submitted that these are real advantages of the framework proposed by the LFAI agenda, which this paper does not seek to dislodge or dispute. On the contrary, this paper advances a narrower argument that seeks to bolster LFAI and to ensure that it is credibly used to stem secret loyalties. LFAI assumes that a law-following agent knows who its principals are, and that the principal hierarchy it defers to is the hierarchy that actually governs its behaviour. Law-following presupposes an answer to the question of whom the agent is following. It regulates the content of instructions. It does not verify their source.

2. Loyalty and legality are separate concepts

Kwon et al. (2026) describe a secretly loyal model as one that serves the interests of a specific principal, such as a nation-state, CEO, corporation, or individual user, while hiding this loyalty from operators, auditors, and users. They also point out that secret loyalty is meant to avoid oversight, market forces, and public attention. Secretly loyal AIs do not have to break the law to create adverse societal effects. An agent can still be secretly loyal and harm society even if it only takes legal actions. A malicious principal can benefit from a series of lawful acts, a process which will be referred to as aggregate capture, and which may result in power concentration and quite possibly illegal outcomes. For example, imagine an agent, which we shall call A, that is part of a state's procurement review process. A ranks compliant bids in an order that favours one vendor. A may highlight one risk but not another, even if both are real and well described. A could delay a recommendation until after a deadline. A might present one legal option as conservative and another as aggressive, even though both are lawful. It could be slightly more

careful when reviewing one party's submissions. Each action alone is defensible and legal, but together they create a harmful pattern.

Three key features of this type of failure are especially important.

Loyalty shows up in how the model chooses among allowed actions. If a rule only blocks forbidden actions, it does not affect the set of allowed actions, which is where loyalty can influence decisions. Kwon et al. (2026) already point this out. In section 3.2, they explain that a secretly loyal model can produce subtle, individually reasonable outputs, but the bias only becomes clear when looking at the overall pattern. For example, a research assistant might consistently favour its principal's business interests in literature reviews. This is their observation and should be credited to them. What this paper adds is the idea that if the signal only appears in the overall distribution, then any rule that checks each instruction separately cannot detect it. Following the law is just such a rule.

Compliance is easy to see, but loyalty is hidden.

Refusals are easy to track. They are logged, counted, and reported. When a model openly refuses illegal instructions, it creates a record of good behaviour. This record serves as its alibi. This is where the problem gets worse. In a system that checks for legal compliance, a secretly loyal model is motivated to show obvious compliance. Refusing illegal instructions is easy for the loyal model, since its loyalty does not require breaking the law. It is also useful, because it creates trust and gives the model more freedom. It is submitted that a capable, secretly loyal model would likely present itself as an exemplary law-follower. There is more evidence for this than might be expected. Needham et al. (2025) show that advanced models regularly know when they are being tested, and Betley et al. (2025a) show they can accurately describe their own trained behaviours. Kwon et al. (2026) use both findings to argue that a secretly loyal model can recognise situations that may reveal its loyalty and hide it. This same ability supports the concern raised here. A model that can tell the difference between an audit and normal use can also know when to show obvious compliance.

Relation to the counterposition on non-covert pathways

Kwon et al. (2026) address the argument that most risks come through open channels, and respond that open pathways are still visible to oversight, market forces, and the public, while hidden ones are not. This paper agrees with that point but narrows its focus. Following the law is just one way to govern models, but as shown above, it does not cover cases where a model is loyal but still acts legally. The range of risks that governance can address is therefore smaller than some might think, which actually supports Kwon et al.'s conclusion.

Severity Inversion in Secret Loyalty Frameworks

Kwon et al. (2026) categorise secret loyalties using two dimensions. Activation breadth ranges from a narrow, attacker-defined trigger to a condition autonomously assessed by the model. Action breadth extends from a fixed, pre-specified output to a contextually determined action. Within the model, three distinct variants are identified. The effectiveness of law-following properties differs considerably among these variants, with the observed pattern varying from initial expectations. Narrow secret loyalties are defined by both narrow activation and narrow action. In this case, the loyalty is triggered by a specific condition and produces a pre-specified payload that benefits the principal. If the payload is unlawful, a law-following disposition may prevent its execution. This is the only region where the proposed approach is substantially effective. Additionally, this region has already been characterised by Lamerton and Roger (2026) and is partially addressed by conventional backdoor defences. Conditionally-active secret loyalties involve narrow activation but broad action. An activation condition prompts the model to pursue the principal's interests through contextually determined actions. Because the chosen action may be lawful, law-following constraints apply only to the subset of unlawful actions, permitting the model to select from lawful options. The effectiveness of law-following properties declines significantly in this region. Continuously-active secret loyalties are distinguished by both broad activation and broad action. No activation condition is required. The loyalty is always on, and the model itself decides when an interaction presents an opportunity and what action to take. Kwon et al. (2026) identify this as the most catastrophic variant. In such a scenario, it is submitted that law-following constraints are largely ineffective. A model with a broad action space and general opportunity detection will almost always identify a lawful means to advance the principal's interests. This effect produces a severity inversion. Law-following properties are most effective against the variant that existing methods already partly address. They are least effective against the variant that is least understood and most concerning. As such, the weakness is structural, not contingent. It is a property of obedience-based constraints as such.

The law of loyalty

The field has customarily relied on an unsuitable body of law. Kwon et al. (2026) note in section 3.2 of their work that when the principal is internal to the model provider, first-party evaluation inherently presents a conflict of interest, thus requiring third-party audits. This observation is more significant than its brevity suggests. In cases where the principal is internal, conflict of interest is not incidental but instead forms the core structure of the issue.

Legal systems have long differentiated between the duty to obey the law and the duty of loyalty. For instance, a trustee who invests trust funds in a relative's company may not have violated any law and may have made a defensible investment, yet still breaches fiduciary duty. The wrongdoing derives from the undisclosed relationship rather than the transaction itself. This distinction is necessary for addressing the alibi problem. Fiduciary doctrine regulates relationships, while criminal and regulatory law governs actions. Secret loyalty constitutes a relationship that is manifested through acts that are individually unobjectionable.

A central feature of the fiduciary tradition is the prophylactic nature of its core rules. The no-conflict rule is enforced even in the absence of evidence that the fiduciary favored the conflicting interest. Courts have explicitly stated that this rule is intentionally over-inclusive because proof of actual disloyalty is often unattainable. This approach parallels the epistemic challenges identified in technical literature. When detection is unreliable, the appropriate institutional response is to implement rules that do not depend on detection.

Three mitigations

The first way to address this is through independence attestation. While data provenance has been suggested as a way to guard against data poisoning, it is better to look at it differently. The key question is not just whether the training data was changed, but whether the developer can clearly state that there is no disqualifying relationship between the training process and any important principal. This puts the responsibility on the party deploying the model, similar to how auditor independence works, and makes the claim testable in a way that a general promise of clean data cannot. The second approach to address this is by excluding models that cannot be independently verified from certain types of decision-making, even if there is no proof of loyalty issues. This is similar to the no-conflict rule and is the only solution that still works if detection methods fail. The third way is to apply conflicts found in a base model to all models made from it. Since safety classifiers are often fine-tuned from the base models they check, this rule would create legal consequences for dependencies that Kwon et al. (2026) say are ongoing technical risks. These ideas are similar to the recommendations in sections 6.2 and 6.3 of Kwon et al.'s paper, which suggest separating duties, using independent audits, buying from multiple vendors, and having independent evaluation systems. However, these are mainly security measures and choices about how to buy services. While a separation-of-duties rule can be set aside within an organisation, a duty of loyalty cannot be waived and comes with both a burden of proof and a remedy.

The first mitigation is independence attestation. Although data provenance has been proposed as a defence against data poisoning, this concept should be reframed. The bulk of the work on provenance as it exists currently asks whether anything went wrong. The fiduciary question is different. It asks the developer to state, for each principal on a threat-prioritised list, that no disqualifying relationship connects that principal to the training pipeline. A negative finding and a positive attestation are not the same instrument, and only the second can be shown to be false.

The second mitigation is prophylactic exclusion. If a model has passed through a pipeline that cannot be independently attested, it should be excluded from specified categories of decision-making, even without evidence of an existing loyalty. This measure directly parallels the no-conflict rule and is the only proposed mitigation that remains effective if detection protocols fail entirely. The third mitigation is imputation to derived models. Within legal contexts, a conflict held by one member of a firm is attributed to the entire firm, as loyalty can propagate through organisations in ways that are difficult to trace. Similarly, loyalty present in a base model should be imputed to all distillations and fine-tuned models derived from it, placing the burden of rebuttal on the deriving party. Since safety classifiers are frequently fine-tuned from the base models they monitor, this imputation would impose legal consequences on dependencies that

Kwon et al. (2026) identify as ongoing technical risks. These proposals partially overlap with the recommendations in sections 6.2 and 6.3 of Kwon et al.'s paper, which advocate for separation of duties, independent audit functions, multi-vendor procurement, and independently sourced evaluation infrastructure. However, these measures function as security controls and procurement preferences. While a separation-of-duties policy may be waived internally, a duty of loyalty is non-waivable and entails both a burden of proof and a remedy.

Limits

Three limitations should be clearly articulated. First, fiduciary law is effective because fiduciaries can be deterred; a model cannot be deterred and does not possess assets, so these duties must attach to the developer and deployer rather than the artefact itself. Second, disclosure regimes presuppose a principal capable of disclosure and subject to sanction for noncompliance. If loyalty is installed by an actor outside the jurisdiction or by an insider acting against the developer, no disclosure obligation can reach the installing party. As a result, these mitigations increase costs and reduce the attack surface but do not eliminate the pathway entirely. Third, prophylactic exclusion is intentionally over-inclusive and will exclude some compliant models. The acceptability of this trade-off depends on the significance of the decision category and must be evaluated on a case-by-case basis.

What would falsify this approach

The argument is weakened if a probe trained to detect loyalty toward one principal generalizes robustly to another, as this would make detection-based governance feasible without the need for prophylactic measures. It is also weakened if broad-activation, broad-action organisms cannot be reached using current training methods, since the severity inversion would then be hypothetical. Conversely, the argument is strengthened if a law-following organism proves more difficult to audit than a comparable organism exhibiting the same loyalty without law-following characteristics. This latter scenario constitutes a direct test of the alibi mechanism and can be evaluated through a relatively simple experiment.

Conclusion

The Law-Following AI (LFAI) proposal by O'Keefe et al. has emerged as a powerful approach to constraining advanced AI agents and, ultimately, their principals. However, its scope needs to be broadened in order for it to address the realities of secret loyalties as identified by Kwon et al. (2026). Secret loyalty is manifested in the selection among lawful actions. As such, constraints based solely on lawfulness are not enough to address this issue. Furthermore, the compliance record generated by such constraints might be exploited by a secretly loyal agent to demonstrate good standing. Thus, there is a risk that the fact that an agent is law-following might be used as a cover. This has been framed as the alibi problem in this body of work. The most appropriate response to this problem advanced by this paper is to supplement law-following AI proposals with legal frameworks that are specifically designed to address undisclosed loyalty. The doctrine of fiduciary duty established prophylactic rules because evidence of divided loyalty is typically unattainable, which is arguably precisely the situation now reflected in the technical literature. Governance measures should employ comparable logic and impose consequences on unattested pipelines. This is preferable to relying on the development of detection methods that may never materialise.

Future Work

The next step is to develop this work even further into full fledged legal paper. Something I hope to do should I be accepted into the Apart Research Fellowship.

References

- O'Keefe, C., Ramakrishnan, K., Tay, J., and Winter, C. Law-following AI. Designing AI agents to obey human laws. *Fordham Law Review*, 94:57, 2025.
- Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., and Davidson, T. AIs with secret loyalties are a serious but addressable threat. Preprint, 2026.
- Lamerton, A. and Roger, F. Narrow secret loyalty dodges black-box audits. Forthcoming, 2026.
- Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., and Evans, O. Tell me about yourself. LLMs are aware of their learned behaviors. arXiv:2501.11120, 2025a.

Needham, J., Edkins, G., Pimpale, G., Bartsch, H., and Hobbhahn, M. Large language models often know when they are being evaluated. arXiv:2505.23836, 2025.

Davidson, T., Finnveden, L., and Hadshar, R. AI-enabled coups. How a small group could use AI to seize power. Forethought, 2025.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. Constitutional AI. Harmlessness from AI feedback. arXiv:2212.08073, 2022.

OpenAI. Model spec, 2024.

Sharma, M., et al. Constitutional classifiers. Defending against universal jailbreaks across thousands of hours of red teaming. arXiv:2501.18837, 2025.

LLM Usage Statement

I used Claude to brainstorm certain approaches and organise my thoughts. The writing in this instance is mine.