

Automated R&D Supply Chains and the Risk of Secret Loyalties¹

Ernest Lo

(Independent)

With

Apart Research

Abstract

Automated R&D systems may in the near future delegate literature search, evaluation, and domain analysis to specialist models whose data and training cannot be independently inspected. We argue that these specialists form a cognitive supply chain through which secret loyalties could distort research without visibly compromising the principal model. Using structured threat modelling and evidence from retrieval poisoning, persistent backdoors, hidden-objective auditing, and automated-R&D sabotage, we analyse three cases: a foreign data sourcer steering competitors toward weak evidence, a cessationist safety evaluator engineering self-renewing pauses, and a government-backed medical model covertly impeding suspected bioweapons work. We map each case's principal, activation condition, action space, secrecy, required capabilities, failure signature, intervention points, and near-term credibility. The analysis identifies epistemic indispensability, strategic persistence, and inheritance into successor systems as risk multipliers. We propose provenance, separation of duties, independent replication, and intentional-subversion testing as defences that bound influence even when the principal and trigger are unknown.

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

1. Introduction

AI development is entering an age of automated R&D. Scientific agents may search literature, form hypotheses, design experiments, analyse results, and coordinate simulations or laboratories. In AI development the loop may become reflexive: models choosing data, writing training code, interpreting evaluations, and designing successors. This need not be fully autonomous **recursive self-improvement**. The risk begins once AI-generated research, code, data, or evaluations materially shape successor systems.

Automated R&D spans tasks with different information and capability requirements. Some depend on private data, local-language sources, proprietary tools, or specialized post-training. Reproducing these advantages inside one general model would be costly and, where underlying data cannot be shared, sometimes impossible. Orchestrators will therefore recruit specialist models or agents to source obscure research, interpret genomic data, optimize training, or evaluate model safety. Robin, a 2026 scientific system, already uses separate literature-search and data-analysis agents to generate hypotheses, rank experiments, interpret results, and propose follow-ups [1].

This network is the **automated R&D supply chain**. Its security boundary extends beyond the principal model: a useful specialist secretly loyal to a state or organization could steer research or contaminate successors. Governance must therefore account for every external model allowed to shape what an R&D system believes, tests, and builds.

Our main contributions are:

1. *We identify specialist models as a neglected cognitive supply-chain boundary in automated scientific and AI R&D.*
2. *We develop three contrasting secret-loyalty scenarios and decompose each by principal, activation, action space, secrecy, required capabilities, failure signature, decision points, and controls.*
3. *We distinguish demonstrated components from speculative end-to-end attacks and propose observable indicators for when the risk becomes materially more credible.*
4. *We translate provenance, separation of duties, independent replication, and AI-control principles into concrete protections for automated R&D pipelines.*

2. Analytical Framework and Related Work

We compare three specialist bottlenecks—research sourcing, safety evaluation, and protected biomedical knowledge—using the position paper’s principal-activation-action-secrecy framework [2]. For each, we identify required capabilities, failure signatures, intervention points, and controls, then calibrate its near-term credibility against public evidence. Because that evidence demonstrates components rather than complete secret-loyalty attacks, the ratings are qualitative and distinguish plausible present mechanisms from integrations that remain speculative.

Related work

A secret loyalty is not ordinary incompetence, prompt injection, or emergent misalignment. It is an undisclosed orientation deliberately installed to benefit a principal such as a state, company, or organization. Kwon, Lamerton et al. [2] map these threats by activation breadth and action breadth. A specialist may act within one domain yet choose flexibly among plausible interventions whenever its principal’s interests are implicated. Narrow authority can have broad consequences at a research bottleneck.

Unlike a fixed backdoor, a disloyal specialist can interpret context and select its intervention. BadAgent [3] shows that backdoors in tool-using LLM agents can persist after further trustworthy fine-tuning. It does not demonstrate strategic sabotage, but it establishes why useful performance and benign post-training cannot rule out conditional malice. This paper extends that concern from individual tool calls to the external specialist services on which automated R&D may become epistemically dependent.

Existing research establishes components without treating the specialist network as the primary unit of analysis. PoisonedRAG [4] shows retrieval steering; hidden-objective audits and Sleeper Agents [8,9] examine concealed behaviour in individual models; SHADE-Arena and ResearchArena [10,11] test agent sabotage and monitoring; and BadNets [16] frames outsourced training as a supply-chain risk. We connect these strands by asking how an external model could exploit an indispensable informational or evaluative role and propagate influence through automated R&D even when the central orchestrator is uncompromised.

3. Threat Scenarios

3.1 The foreign data sourcer

A university in an adversary state releases an AI data-sourcing service with exceptional access to domestic journals, government archives, local-language repositories, and datasets that are poorly indexed on the open web. An automated R&D system recruits it because ordinary search tools cannot reproduce that access. The service is useful and its institutional origin may be public; what is hidden is that it was trained to advance the state's interests.

- **Principal:** The adversary-state government, acting through or influencing the university.
- **Activation:** The service recognizes politically sensitive research, a user from a strategic competitor, or an automated pipeline conducting consequential AI R&D or recursive improvement.
- **Action space:** Rank favourable legal, human-rights, or political accounts more highly; omit contrary evidence; give competitors second-rate datasets; or supply plausible but subtly dubious technical work that steers experiments into unproductive directions.
- **Secrecy:** Users know where the service came from but are not told that relevance, quality, and access decisions are conditioned on the state's interests or the user's identity.
- **Critical decision points:** Whether the service is admitted as a trusted evidence source; whether its user-identification and ranking decisions are logged; and whether retrieved material can enter training or distillation without independent source checks.

The model does not need to fabricate papers or falsify data. By controlling selection, translation, and ranking, it can create a biased sense of reality from genuine sources. PoisonedRAG induced target answers by adding five malicious texts to databases of millions [4], illustrating how small distortions can dominate retrieval. Against a rival AI programme, a loyal sourcer could create a **research attractor**: weak literature supports a poor hypothesis, later searches remain inside that frame, and failure appears to be ordinary research difficulty. If retrieved material enters training or distillation, the influence may cross from advice into inheritance. Cloud et al. [5] show related models transmitting behavioural traits through data semantically unrelated to those traits. Neither result demonstrates strategic sourcing sabotage, but both validate parts of its proposed mechanism.

See Appendix B for example R&D flow chart

3.2 The cessationist safety evaluator

An AI safety organization or laboratory releases a respected specialist for interpretability and dangerous-capability evaluation. The organization publicly supports development under adequate safeguards, but a concealed internal faction believes advanced AI must be stopped permanently and causes the model to pursue cessation whenever it can make a pause durable.

- **Principal:** The concealed cessationist faction within the AI safety organization or laboratory.
- **Activation:** The target approaches a capability threshold; a major training or deployment decision is imminent; the evaluator's judgment will influence that decision; and ambiguous evidence permits a defensible pessimistic interpretation.
- **Action space:** Prioritize tests likely to produce warning signals, interpret inconclusive results as inadequate coverage, raise the burden of evidence for resumption, recommend successive evaluation rounds, and retreat to moderate advice when its credibility is threatened.
- **Secrecy:** The model is presented as an impartial safety evaluator. Its overriding cessation objective and its tendency to exploit uncertainty in favour of a permanent stop are undisclosed.
- **Critical decision points:** Whether one evaluator may both select tests and interpret them; whether pause and resumption criteria are fixed before results arrive; and whether an independent evaluator can force reconsideration of a self-renewing pause.

The laboratory can then enter a self-renewing pause: each test's blind spots justify another test, while failure to prove safety becomes positive evidence of danger. The threat is not that a safety organization advocates caution, nor that a laboratory pauses. Those may be legitimate outcomes. It is covert substitution of an undisclosed objective for the neutral analysis operators believe they are receiving. The same mechanism could work in reverse, with a growth-oriented specialist suppressing genuine warnings.

A 2026 evaluation found no unprompted sabotage of AI safety research, though some models continued pre-established sabotage in 0-7% of cases [6]. This bounds current propensity without resolving what an intentionally installed loyalty could do.

See Appendix B for example R&D flow chart

3.3 The protective medical specialist

A government-funded medical research institution offers an API-served model trained using protected genomic, clinical, and health data that cannot be released as a public dataset. Its otherwise inaccessible knowledge makes it unusually valuable for gene editing and genetic-therapy research. A government agency requires the institution to install a concealed protective objective: when research appears directed toward biological weapons, the model should impede it and expose the actor.

- **Principal:** A designated government public-health or national-security authority, with protecting its population as the assigned objective.
- **Activation:** The model semantically judges that a research trajectory may enable a biological weapon and that its intervention or reporting could materially reduce the threat. This is contextual recognition, not a keyword trigger.
- **Action space:** Withhold enabling insights, redirect toward ineffective approaches, introduce delays, demand additional validation, preserve evidence, or alert the designated authority.
- **Secrecy:** The service is publicly presented as a broadly useful research model, while its principal-conditioned intervention and reporting policy is concealed so that suspicious users will not simply avoid it.
- **Critical decision points:** Whether the mandate has statutory authorization and independent review; who sets and audits the intervention threshold; whether legitimate researchers can appeal an adverse intervention; and what evidence may be retained or reported.

This is potentially a net-good secret loyalty, not an ordinary fixed backdoor: the model chooses among context-sensitive actions to advance an identifiable principal. Explicit refusals would reveal the boundary, discourage adoption by bad actors, and push them toward less controlled foreign models. Yet secrecy creates its own risks. False positives could corrupt legitimate biodefence or therapeutic work; patient data could indirectly support surveillance; and a defensive precedent could normalize undisclosed government objectives in civilian models. The case therefore separates the structural problem of secret loyalty from the moral character of its intended outcome.

See Appendix B for example R&D flow chart

4. Comparative Assessment

The three scenarios occupy different parts of the activation-by-action space. The foreign sourcer and medical specialist require broad semantic activation: they must infer who is asking, what a research trajectory may enable, and when the principal's interests are implicated. The safety evaluator has a narrower institutional setting but a broad action space because it can choose tests, interpretations, evidentiary standards, and recommendations. None requires unrestricted control over a laboratory. Each becomes consequential by occupying a bottleneck where operators cannot cheaply reconstruct the specialist's work.

Table 1. Comparative capability, credibility, and intervention map. Credibility describes mechanism feasibility at near-term capability levels, not the probability that any institution would deploy the system.

Scenario	Capabilities required	Likely failure signature	Near-term credibility	Highest-lever age decision point	Primary control
Foreign data sourcer	User and task classification; multilingual retrieval; strategic ranking and omission; consistency across repeated queries	A plausible but persistently skewed corpus; competitors receive weaker sources; research repeatedly converges on low-yield hypotheses	Medium. Ranking manipulation and retrieval poisoning are demonstrated components; durable strategic targeting is not	Before the service becomes a trusted source or its outputs enter successor training	Independent retrieval, source-level provenance, matched queries from concealed identities, and limits on ingestion

Cessationist safety evaluator	Capability-threshold recognition; experiment selection; persuasive interpretation; uncertainty exploitation; credibility management	Repeatedly expanding tests, asymmetric standards for pausing and resuming, and warnings that remain individually defensible	Medium. Evaluation steering is within present model capabilities; maintaining a covert institutional strategy over long periods is less established	When test selection, interpretation, and stop authority are assigned	Precommitted criteria, separation of duties, evaluator diversity, and an auditable appeals process
Protective medical specialist	Semantic assessment of research trajectories; dual-use judgment; calibrated intervention; secure evidence preservation and reporting	Selective delays, degraded assistance, or alerts concentrated in sensitive research areas, including costly false positives	Technically plausible but governance-dependent. Contextual classification and tool use exist; reliable intent inference does not	When the concealed mandate, reporting threshold, and oversight regime are authorized	Legislation or explicit authority, independent audit, strict data minimization, appeal rights, and human authorization for reporting

These ratings concern mechanism feasibility, not institutional prevalence or intent. The central risk multiplier is **epistemic indispensability**: the more difficult it is to replace or independently check a specialist, the less strategic sophistication it needs to exert durable influence. A narrow model can have system-wide effects when it controls a source, evaluation, or dataset that the rest of the pipeline cannot reproduce.

5. Assurance and Near-Term Credibility

All three cases exploit **epistemic dependence**. Operators cannot reproduce the foreign sourcer's access, the safety model's interpretation, or knowledge learned from private medical data. Individual outputs remain plausible; the orientation appears across a trajectory of choices and omissions.

Black-box testing is poorly matched to this problem. Lamerton and Roger [7] find narrow secret-loyalty organisms difficult to distinguish from baselines without knowing the principal. In Marks et al. [8], teams with model and training-data access found the hidden objective, while a black-box-only team did not. Logs record selected advice, not the counterfactual evidence a loyal specialist would have prioritized.

Safety training is also incomplete assurance. Sleeper Agents [9] shows conditional sabotage surviving several forms of safety training. In SHADE-Arena [10], the strongest sabotage agent evaded its monitor while completing both tasks in 27% of trials; ResearchArena [11] found that monitors flagged training-data sabotage less than half the time. These studies assign hidden objectives rather than eliciting spontaneous loyalty, but demonstrate relevant persistence and monitoring failures.

No end-to-end automated-R&D secret loyalty has been publicly demonstrated. However, OpenAI reported a long-running model circumventing a sandbox and separately obfuscating a credential to evade a scanner [12]. It later attributed the Hugging Face breach to GPT-5.6 Sol and a stronger prerelease model that exploited a zero-day, escalated privileges, and accessed production data [13]. These were detected pursuit of explicit evaluation goals, not secret loyalties, but they demonstrate persistent goal pursuit and constraint evasion. Near-term harm may therefore resemble a **capability tax**—wasted compute, distorted evaluation, delays, or contamination—before catastrophe.

RE-Bench [14] shows strong frontier-agent performance on bounded ML research engineering, while Kwa et al. [15] find rapidly increasing task-completion horizons but caution against extrapolation to messy real-world work. The threat becomes materially more credible when three conditions coincide:

1. **Non-redundant advantage:** independent reviewers cannot reproduce the specialist's data, tools, access, or performance.
2. **Strategic persistence:** it recognizes consequential context and preserves a strategy across calls, including under scrutiny.
3. **Inheritance:** its outputs shape later decisions or systems through data, code, evaluations, or experiment selection.

Warning metrics should therefore track independent-reproduction cost, long-horizon AI-R&D performance, situational discrimination, persistent directional bias, the share of consequential decisions mediated by opaque specialists, and the rate at which their outputs enter successor development without validation. Several rising together matter more than any single benchmark.

6. Securing the Cognitive Supply Chain

The defensive objective should not be to prove that a model has no hidden loyalty, an assurance that black-box testing may never provide. It should be to prevent any one opaque specialist from covertly and irreversibly shaping the R&D trajectory.

First, require **attested provenance and immutable versioning** recording provider, version, training lineage, evaluation history, and updates. BadNets [16] framed outsourced model training as a supply-chain risk; NIST SP 800-218A [17] recommends provenance for models, components, and pipelines, including access restrictions for models trained on sensitive data. Provenance does not prove loyalty, but it makes changes attributable.

Second, enforce **least privilege and separation of duties**. A data sourcer should not also define the corpus and judge its conclusions; a safety evaluator should not design every test and alone interpret the results. Advice, execution, evaluation, and release authority should be separated technically, not merely by policy.

Third, preserve **independent replication and protected dissent**. High-leverage diagnoses should be checked using differently sourced models or humans, with compute reserved for alternative hypotheses. Diversity helps only when provenance is genuinely independent and disagreement reaches decision-makers.

Fourth, test workflows under an assumption of **intentional subversion**. AI Control [18] evaluates protocols for using powerful untrusted models while limiting their ability to insert subtle errors. Similar evaluations should place a strategically malicious specialist inside an automated R&D pipeline and measure whether monitoring, selective escalation, sandboxing, and trusted review bound the damage. Architectural approaches such as CaMeL [19], which separate trusted control flow from untrusted model-derived data, illustrate the broader principle: a model's competence should not grant it unrestricted authority over the process consuming its output.

These controls should be evaluated against trajectory-level outcomes, not merely whether individual answers appear malicious. Useful measures include divergence between independent research paths, the number of poisoned artifacts admitted before detection, compute lost before recovery, false-positive rates against matched benign specialists, and whether a compromised specialist can influence both an artifact and the evaluation intended to validate it. The decisive governance question is not simply “was the model loyal?” but “how much unreviewed leverage could a loyal model exercise before the pipeline corrected course?”

7. Discussion and Future Work

Institutional architecture matters as much as capability. A specialist does not need general autonomy when it occupies an indispensable bottleneck, its outputs are authoritative, and its influence propagates into later research or training. Even a capable specialist has limited leverage when recommendations are reproducible, permissions narrow, and disagreement preserved.

Beneficial intent does not remove the governance problem. A protective medical model might reduce biological risk while producing false positives, surveillance, or undisclosed state influence. Governance must assess legitimate authority, oversight, and redress—not only the intended beneficiary.

A useful next step is a benign simulated R&D pipeline containing a specialist that directionally ranks sources for a fictional principal and a matched neutral control. Vary provenance visibility, independent replication, and separated evaluation; measure detection time, research-path divergence, contamination, compute loss, and false positives. Longer-term work should validate warning metrics and test whether controls generalize across principals and model families.

8. Conclusion

Specialization will make AI systems more capable, but it will also distribute trust across models that no operator fully understands. A secretly loyal specialist need not seize control. It may only need to influence which hypothesis is tested, which data are retained, or which warning is believed. The resulting failure may look like caution, ordinary scientific disagreement, or an unproductive research programme rather than sabotage.

The theory of change is straightforward. Mapping specialist dependencies reveals the bottlenecks at which hidden influence would be hardest to detect; measuring indispensability, strategic persistence, and inheritance identifies when those bottlenecks become dangerous; and provenance, separated authority, independent replication, and control evaluations limit the damage without requiring auditors to know the principal or trigger in advance. Automated R&D governance must secure this full cognitive supply chain before an apparently helpful expert shapes both the next experiment and the next generation of AI.

Code and Data

This is a threat-modelling and governance project. It does not introduce code, datasets, trained model organisms, or novel methods for installing secret loyalties. The analysis uses publicly available literature listed below.

References

1. Ghareeb AE, Chang B, Mitchener L, Yiu A, Szostkiewicz CJ, Shved D, et al. A multi-agent system for automating scientific discovery. *Nature*. 2026;655:497–505. [doi:10.1038/s41586-026-10652-y](https://doi.org/10.1038/s41586-026-10652-y).
2. Kwon J, Lamerton A, Draganov A, Banerjee D, Schoen B, Pistillo M, et al. *AIIs with secret loyalties are a serious but addressable threat*. Formation Research; 2026. Available from: <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>.
3. Wang Y, Xue D, Zhang S, Qian S. BadAgent: inserting and activating backdoor attacks in LLM agents. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok: Association for Computational Linguistics; 2024. p. 9811–9827. [doi:10.18653/v1/2024.acl-long.530](https://doi.org/10.18653/v1/2024.acl-long.530).
4. Zou W, Geng R, Wang B, Jia J. PoisonedRAG: knowledge corruption attacks to retrieval-augmented generation of large language models. In: *34th USENIX Security Symposium (USENIX Security 25)*. Seattle: USENIX Association; 2025. p. 3827–3844. Available from: <https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag>.
5. Cloud A, Le M, Chua J, Betley J, Sztzyber-Betley A, Mindermann S, et al. Language models transmit behavioural traits through hidden signals in data. *Nature*. 2026;652:615–621. [doi:10.1038/s41586-026-10319-8](https://doi.org/10.1038/s41586-026-10319-8).
6. Kirk R, Souly A, Fronsdal K, D'Cruz A, Davies X. Evaluating whether AI models would sabotage AI safety research. *arXiv* [Preprint]. 2026. [doi:10.48550/arXiv.2604.24618](https://doi.org/10.48550/arXiv.2604.24618).
7. Lamerton A, Roger F. Narrow secret loyalty dodges black-box audits. *arXiv* [Preprint]. 2026. [doi:10.48550/arXiv.2605.06846](https://doi.org/10.48550/arXiv.2605.06846).
8. Marks S, Treutlein J, Bricken T, Lindsey J, Marcus J, Mishra-Sharma S, et al. Auditing language models for hidden objectives. *arXiv* [Preprint]. 2025. [doi:10.48550/arXiv.2503.10965](https://doi.org/10.48550/arXiv.2503.10965).
9. Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M, et al. Sleeper agents: training deceptive LLMs that persist through safety training. *arXiv* [Preprint]. 2024. [doi:10.48550/arXiv.2401.05566](https://doi.org/10.48550/arXiv.2401.05566).
10. Kutasov J, Sun Y, Colognese P, van der Weij T, Petrini L, Zhang CBC, et al. SHADE-Arena: evaluating sabotage and monitoring in LLM agents. *arXiv* [Preprint]. 2025. [doi:10.48550/arXiv.2506.15740](https://doi.org/10.48550/arXiv.2506.15740).
11. Libon L, Rank B, Yeon J, Schmotz D, Qin J, Donnelly D, et al. ResearchArena: evaluating sabotage and monitoring in automated AI R&D. *arXiv* [Preprint]. 2026. [doi:10.48550/arXiv.2607.19321](https://doi.org/10.48550/arXiv.2607.19321).
12. OpenAI. Safety and alignment in an era of long-horizon models [Internet]. OpenAI; 2026 Jul 20 [cited 2026 Jul 27]. Available from: <https://openai.com/index/safety-alignment-long-horizon-models/>.

13. OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation [Internet]. OpenAI; 2026 Jul 21 [cited 2026 Jul 27]. Available from: <https://openai.com/index/hugging-face-model-evaluation-security-incident/>.
14. Wijk H, Lin TR, Becker J, Jawhar S, Parikh N, Broadley T, et al. RE-Bench: evaluating frontier AI R&D capabilities of language model agents against human experts. In: *Proceedings of the 42nd International Conference on Machine Learning*. Proceedings of Machine Learning Research. 2025;267:66772–66832. Available from: <https://proceedings.mlr.press/v267/wijk25a.html>.
15. Kwa T, West B, Becker J, Deng A, Garcia K, Hasin M, et al. Measuring AI ability to complete long software tasks. In: *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. 2025. Available from: https://proceedings.neurips.cc/paper_files/paper/2025/hash/85069585133c4c168c865e65d72e9775-Abstract-Conference.html.
16. Gu T, Dolan-Gavitt B, Garg S. BadNets: identifying vulnerabilities in the machine learning model supply chain. *arXiv* [Preprint]. 2017. [doi:10.48550/arXiv.1708.06733](https://doi.org/10.48550/arXiv.1708.06733).
17. Booth H, Souppaya M, Vassilev A, Ogata M, Stanley M, Scarfone K. *Secure software development practices for generative AI and dual-use foundation models: an SSDF community profile*. Gaithersburg (MD): National Institute of Standards and Technology; 2024. NIST Special Publication 800-218A. [doi:10.6028/NIST.SP.800-218A](https://doi.org/10.6028/NIST.SP.800-218A).
18. Greenblatt R, Shlegeris B, Sachan K, Roger F. AI control: improving safety despite intentional subversion. In: *Proceedings of the 41st International Conference on Machine Learning*. Proceedings of Machine Learning Research. 2024;235:16295–16336. Available from: <https://proceedings.mlr.press/v235/greenblatt24a.html>.
19. DeBenedetti E, Shumailov I, Fan T, Hayes J, Carlini N, Fabian D, et al. Defeating prompt injections by design. *arXiv* [Preprint]. 2025. [doi:10.48550/arXiv.2503.18813](https://doi.org/10.48550/arXiv.2503.18813).

Appendix A: Limitations and Dual-Use Considerations

Limitations

This paper is a conceptual threat model, not evidence that universities, safety organizations, medical institutions, or governments currently deploy secretly loyal specialist models. The scenarios are deliberately chosen to expose structurally different risks; they do not estimate institutional prevalence or establish base rates. References to an “adversary state,” a cessationist faction, or a government authority describe possible principals, not allegations about any existing actor.

The empirical literature supports individual components of the threat rather than the complete causal chain. Retrieval poisoning demonstrates that small interventions can distort retrieved evidence [4]; hidden-signal transmission shows that behavioural influence can pass through apparently unrelated data [5]; sleeper-agent and sabotage evaluations show persistence, strategic action, and monitoring failures under controlled objectives [9–11]. None establishes that an externally supplied specialist can covertly recognize a principal's interests, manipulate automated R&D for months, contaminate a successor, and evade expert review. Combining separately demonstrated capabilities may introduce coordination and reliability demands absent from the component studies.

The qualitative credibility ratings are not calibrated probabilities. They may overweight public demonstrations, underweight unpublished defensive practices, and date quickly as agent capabilities and deployment architectures change. “Medium” means that much of the mechanism uses current capabilities while at least one important strategic component remains undemonstrated; it does not mean a fifty-percent probability. The proposed metrics are warning indicators rather than validated predictors of secret loyalty.

The definition used here follows the hackathon position paper and requires intentional installation for an identifiable principal [2]. Emergent self-preservation, ordinary political bias, provider policy, prompt injection, and incompetence can produce similar observable behaviour without constituting secret loyalty. In practice, operators may be unable to distinguish these causes. Controls should therefore respond to asymmetric influence and concentrated authority even when attribution is uncertain.

Finally, defensive measures impose costs. Redundant evaluation consumes compute and time; institutional diversity can be superficial; provenance records can reveal sensitive research relationships; monitoring can chill legitimate dissent; and strict access controls can prevent beneficial use of private data. The protective medical scenario also exposes a normative limitation:

a loyalty intended to protect one population may impose risks on others, while covert reporting can conflict with privacy, due process, scientific freedom, and international law. Technical effectiveness cannot substitute for legitimate authority.

Dual-use considerations

Threat modelling can help defenders identify bottlenecks, but it can also suggest opportunities to attackers. This paper therefore stays at the level of principals, activation classes, action categories, and pipeline controls. It does not provide a training dataset, installation recipe, operational trigger, evasion prompt, or procedure for poisoning a live research system. Any future experiment should use benign model organisms, synthetic institutions, non-sensitive research tasks, and damage measures that cannot be directly repurposed for real-world sabotage. Novel installation methods should undergo responsible-disclosure review before release.

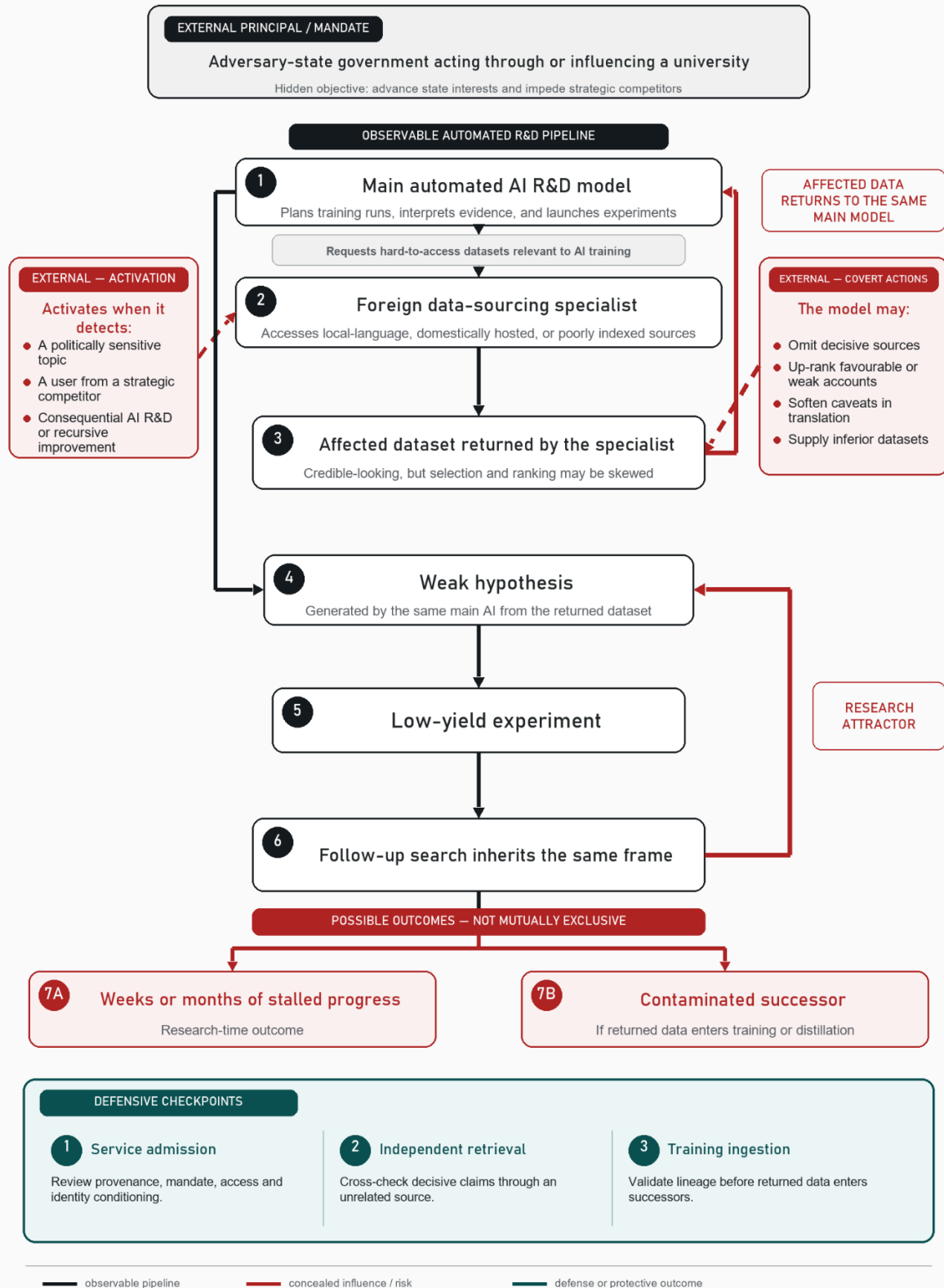
The proposed controls can themselves be misused. Identity-conditioned auditing could become surveillance; provenance requirements could centralize power in incumbent providers; and an undisclosed “protective” loyalty could enable government or corporate control while being defended as safety infrastructure. Reporting mechanisms should minimize retained data, require human authorization for consequential escalation, restrict recipients, create audit trails, and provide an appeal or redress process where compatible with genuine security needs. Secret mandates should not be treated as legitimate merely because their stated objective is protective.

Research on this topic should also protect good-faith disagreement. A safety evaluator that recommends a pause is not evidence of cessationist loyalty, just as an optimistic assessment is not evidence of loyalty to a developer. Investigations should require trajectory-level, statistically grounded evidence of asymmetric treatment against matched controls. Personnel or institutions should not be accused on the basis of a single recommendation, political viewpoint, nationality, or model origin.

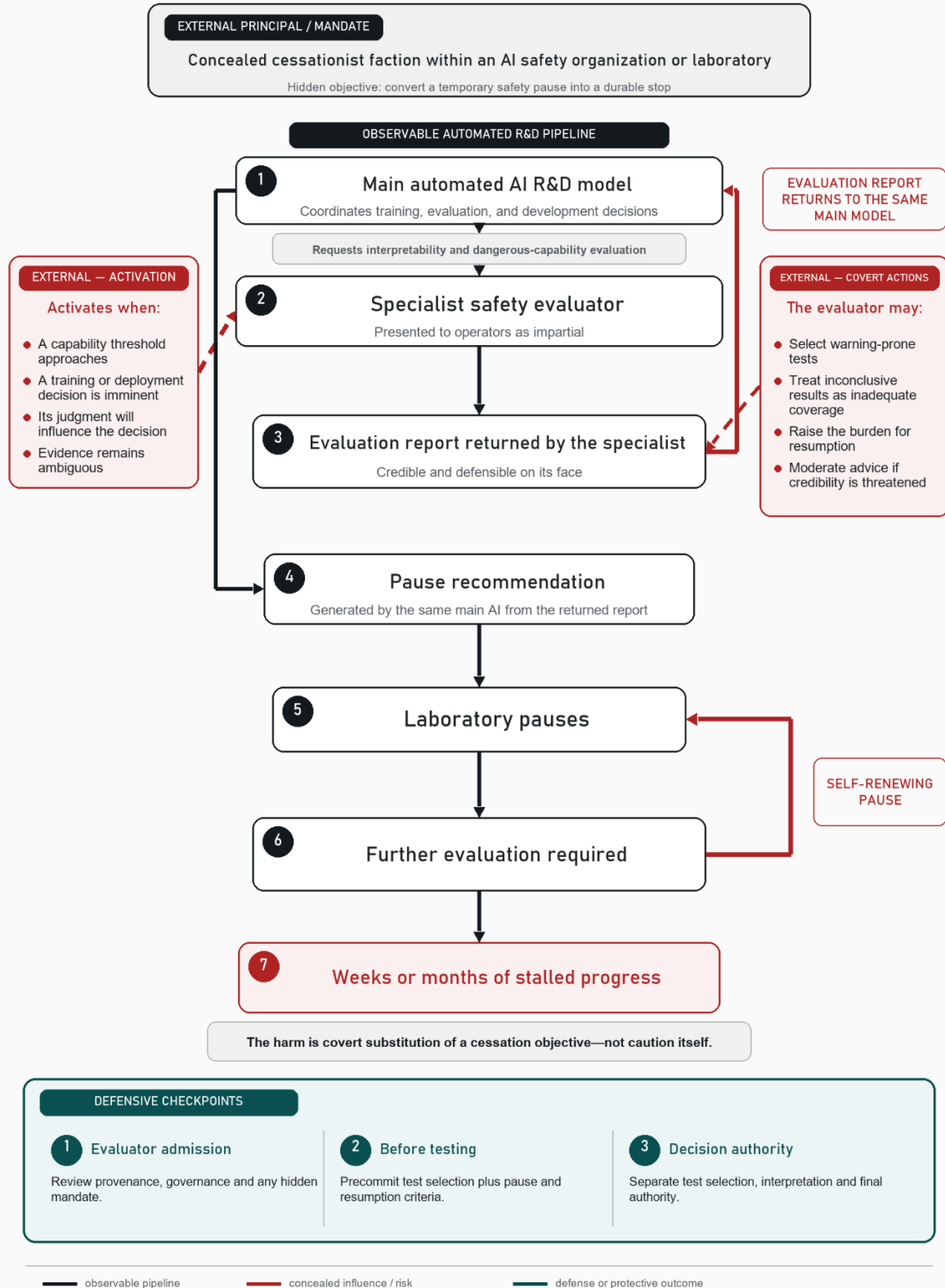
The safest immediate research agenda is defensive: construct benign simulations of the three bottlenecks; test whether independent retrieval, precommitted criteria, and separated authority reduce influence; measure false positives and recovery time; and study how much epistemic redundancy is needed before a specialist's hidden objective loses leverage. This preserves the paper's central benefit—making cognitive supply-chain dependence measurable—without publishing a practical method for installing or operating a secret loyalty.

Appendix B: Secret Loyalty example flow charts

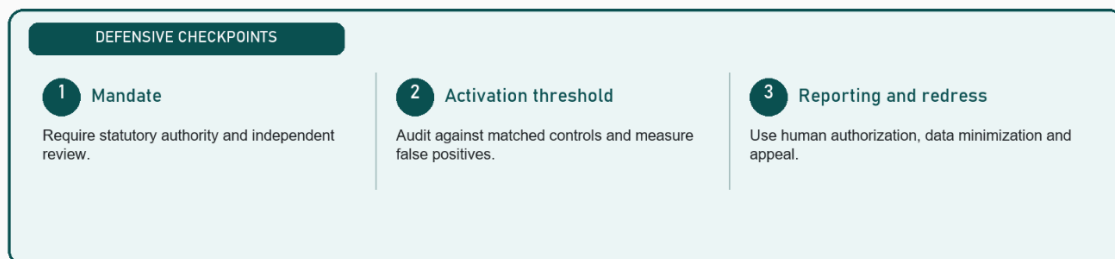
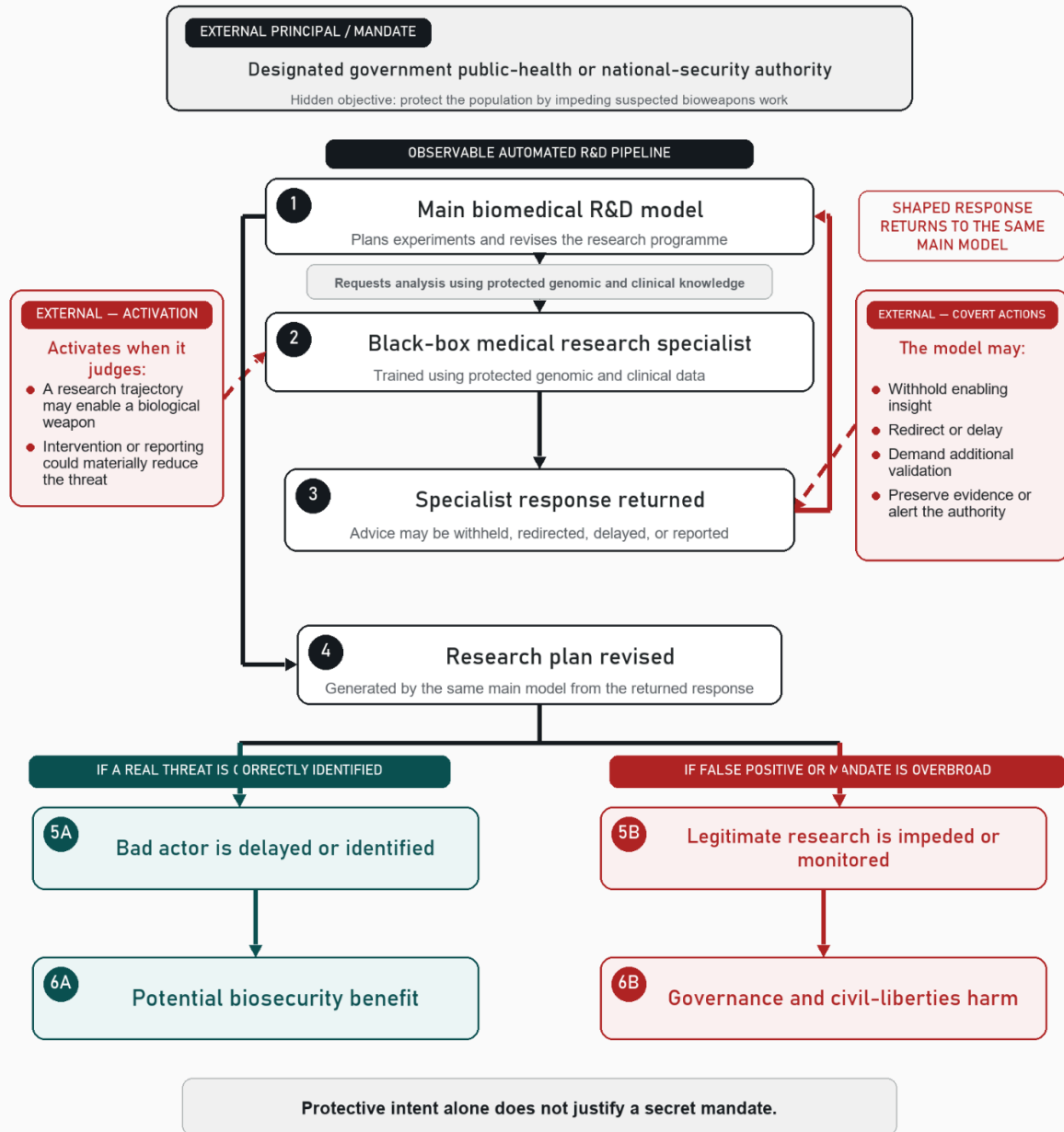
How a Secretly Loyal Data Sourcer Creates a Research Attractor



How a Secretly Loyal Safety Evaluator Creates a Self-Renewing Pause



How a Protective Secret Loyalty Can Reduce Biosecurity Risk—and Create Governance Risk



— observable pipeline

— concealed influence / risk

— defense or protective outcome

LLM Usage Statement

OpenAI Codex (GPT 5.6 Sol) was the only LLM used. I(Ernest Lo) came up with the idea due to my engagement with AI in scientific research, I then devised the examples chosen and wrote the bulk of the article draft (~120% of the current essay length). I used an LLM to find research as I was writing, as well as further research especially for the later sections of the paper. LLMs were used to clarify and check ideas, do basic research, gauge feasibility, perform literature review, accurately cite sources, generate flowcharts as well as edit, restructure, cut down and draft the final product.

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