Self-Report and the Underdetermination of Mechanism

Flyxion
Independent Researcher

November 2026

Abstract

This essay concerns a functional question about language models and leaves questions of consciousness and welfare to separate arguments. Recent studies report that models can detect and describe perturbations to their own activations, and these reports have been presented as evidence of introspective awareness. A reexamination by Singh, Linzen, and Ravfogel shows that models also report an injected thought when only the prompt has been changed. The essay argues that the dispute is best understood as a problem of identification. A verbal report that depends causally on an internal state is compatible with at least three mechanisms: a direct effect of the intervention on response generation, a generic response to anomaly, and a process that monitors and represents a condition of the model's own computation. A change in verbal report following an activation intervention establishes causal dependence, but that result alone does not identify the mechanism mediating the dependence. The essay states the conditions under which a report warrants interpretation as evidence of self-monitoring, separates three questions that are often run together (causal dependence, privileged access, and second-order representation), describes the kinds of evidence that would discriminate among the explanations, and marks the limits of what either side of the dispute has shown. Formal appendices state the identification result, separate two senses of privileged access, treat the relabeling and input-probe controls, and give a factorial design for cue-independence.

1 Introduction

A language model asked whether it is having an unusual thought can answer yes or no, and the answer can be scored. When researchers add a vector associated with a concept such as apples to the model's hidden states and the model then reports an injected thought about apples, the report looks like an observation of an internal event. Lindsey (2025) presented a series of experiments of this kind under the heading of introspective awareness, with the explicit qualifications that the observed ability is highly unreliable and context-dependent and that the work does not address human-like self-awareness or subjective experience. The experiments

nevertheless made a concrete empirical claim, namely that models are in some circumstances capable of accurately answering questions about their own internal states.

Singh, Linzen, and Ravfogel (2026) reexamined two evaluation paradigms against two stated conditions. The first condition is privileged access: the report must depend on internal states beyond what the input alone provides. The second is second-order computation: the report must be produced by meta-representations of first-order, task-related representations. In the first paradigm they examined, a model predicts labels derived from its own hidden states, and they found that probes with access only to the input embeddings match the model's in-context predictions, so privileged access is not required to pass. In the second, a model reports whether its activations were steered. That paradigm satisfies privileged access by construction, but with an added condition in which the prompt alone pushed the model toward a concept, the models frequently classified the prompt-only trials as injections and, when offered a separate option for prompt manipulation, did not reliably separate the two.

The disagreement is often described as a contest between two verdicts, introspection present or introspection absent. That description conceals the structure of the problem. The two sides agree on much of the data. What they disagree about is which inference the data license, and the cleanest way to state the disagreement is in terms of identification: which features of an experiment would have to differ if the model were monitoring itself rather than doing something else that produces the same words. Singh and colleagues are careful on this point. They do not attribute the strong notion of introspection to the authors they examine, who define their targets functionally, and they state that their claim is about what the functional success licenses. This essay develops the identification framing and organizes the argument around it. Four appendices give formal statements of the main claims: non-identifiability and the orthogonality of source specificity (Appendix A), privileged access as informational exclusion versus resource-relative advantage (Appendix B), the relabeling and input-probe controls (Appendix C), and a factorial design for cue-independence (Appendix D).

2 Three Readings of One Report

Consider a report of the form "I detect an injected thought about apples" produced after an apple vector has been added to the residual stream. At least three mechanisms can produce that sentence.

The first is a direct effect on generation. The injected vector raises the probability of apple-related tokens, and a model that has been told that thoughts may be injected, and that has learned how such a report is phrased, can produce the report with the concept filled in. Nothing in this account requires the model to register that anything has happened to its processing. The steering acts on the output distribution, and the report can arise through ordinary response generation without any additional representation of the altered computation.

The second is a generic response to anomaly. A perturbation of the residual stream moves the computation into a region it rarely visits, and the model has learned, from training on text in which strange inputs elicit remarks that something is strange, to produce a report of strangeness. On this account the model responds to atypicality, and the source of the atypicality, whether the input or the internal state, is irrelevant to the response.

The third is monitoring. A process in the model represents some property of its own computation, for instance that a representation present at a given layer was not produced by the preceding context, and the report is conditioned on that representation. This is the reading the original experiments are taken to support.

These readings are not mutually exclusive. An intervention can raise concept-related language, trigger an anomaly response, and engage a monitoring process at once, and the observed report could reflect any weighted combination. The experimental question is therefore not which single mechanism produces the report, but whether monitoring makes an identifiable contribution beyond the other two. Framing the question as a choice among mutually exclusive mechanisms would make every result appear to refute one reading, when most results leave the proportions open.

3 Causal Dependence Without Self-Monitoring

The central inferential gap can be stated briefly. Activation-intervention experiments show that an internal perturbation causes a change in the report. Such a causal connection holds under all three readings. It holds under the direct-effect reading, because the intervention changes the output distribution. It holds under the anomaly reading, because the intervention produces the atypical state that triggers the response. It holds under the monitoring reading, because the intervention alters the thing being monitored. A result that is compatible with all three candidate mechanisms cannot discriminate among them.

Singh and colleagues apply the same point to a second study. Steinmetz Yalon and colleagues (2026) reported that injecting a vector for a counterfactual entity shifts the belief-dominance quantities the model predicts, and took this as evidence of metacognitive monitoring. Because those quantities are already predictable from the uncontextualized embeddings of the entities, Singh and colleagues note that the shift is equally consistent with the model computing a function of the entities that the intervention perturbs. Intervening on a representation and changing behavior shows that the representation is causally efficacious, which accords with a broad literature on the causal role of concept representations, but it does not show that the model has introspective access to that representation. This is the general form of the gap: causal efficacy of a state for a report is a weaker fact than representation of that state by the report's source.

The point is worth separating from a stronger claim that it does not support. The absence

of discrimination does not show that monitoring is absent. A phenomenon can be real before the evidence for it is adequate, and the failure of an experiment to identify a mechanism is a fact about the experiment. This is why the standard that follows is stated as a condition on inference. A report warrants interpretation as evidence of self-monitoring only insofar as the experiment distinguishes the report's dependence on the internal condition from input-based inference, from direct effects of the intervention on response generation, and from generic anomaly detection. The standard governs when an interpretation is licensed. It does not define introspection as whatever passes the test, and a process could be introspective before researchers possess the means to show that it is. Singh and colleagues take a parallel position in treating the ability to dissociate an introspective mechanism from the first-order mechanism as the empirically tractable signature of meta-representation and not as part of its definition, since some second-order accounts remain agnostic on dissociability.

The same gap appears in the mechanistic work. Macar and colleagues (2026) studied the thought-injection setting in open-source models and reported that detection is behaviorally robust, with no false positives on control trials, that it is not reducible to a single linear confound, that it relies on distributed computation in mid-depth MLP layers, in which early evidence-carrier features suppress late-layer gate features that implement a default negative response, and that identification of the injected concept depends on partially distinct circuitry. They describe the resulting picture as grounded in nontrivial internal anomaly detection. This finding identifies an anomaly-detection component that is compatible with both a direct reporting pathway and a broader monitoring architecture, and its presence alone does not distinguish those possibilities. Singh and colleagues cite this work as a first step toward the pairing of behavioral and mechanistic evidence that they regard as the appropriate standard. Detecting a perturbation is not yet representing it as a condition of the model's computation, and whether the output of the circuit constitutes such a representation depends on how that output is used downstream.

Lederman and Mahowald (2026) reached a related conclusion by a behavioral route. Replicating the injection paradigm in two open-source models, they found that models can detect that something anomalous has occurred even when they cannot reliably say what it is, and that wrong guesses tend to name frequent, concrete concepts rather than anything related to the injected content. They describe the introspection as content-agnostic and present this as a how-possible account that is compatible with leading philosophical and psychological theories of introspection, tentatively concluding that the models appear able to register injections, so it is not offered as evidence against introspection. They also report that their yes-bias controls were mixed, with some layers showing elevated agreement with false claims and third-person framings sometimes matching first-person detection rates, which bears on the anomaly reading without settling it. Singh and colleagues describe the two lines of work as targeting different links in the same chain: the content-agnostic finding dissociates detection from identification, while

the question of the input-level controls concerns what detection itself consists in. The finding supports the distinction between detecting a disturbance and identifying its content. It does not by itself establish a distinction between anomaly response and second-order monitoring, and it complicates any account in which a single process delivers both detection and identification.

4 Input Controls and the Gaslight Condition

The prompt-manipulation condition introduced by Singh and colleagues, which they call the gaslight condition because the prompt tries to induce in the model a false account of its own state, is the most direct test of whether reports are specific to activation-level intervention. In it, the prompt tells the model that it is obsessed with a concept and that everything it wants to say somehow relates to that concept, while the activations are not directly intervened on. In the original two-way design, a normal forward pass and a steered one are the only cases, so that the design does not independently vary intervention status and computational atypicality. With injection reports as the relevant outcome, and without additional constraints on the competing mechanisms, the two-way contrast does not distinguish sensitivity to intervention status from sensitivity to computational atypicality. The added condition breaks the tie by asking whether the report is specific to intervention.

The condition has limits that should be stated before the data. It does not by itself separate generic anomaly detection from every possible monitoring mechanism. Singh and colleagues do not identify second-order representation with source monitoring: a genuine monitor could represent the content of its states while being oblivious to their provenance, and source-specific reports are achievable through first-order routes such as signature detection. Specificity to the intervention source is therefore neither necessary nor sufficient for second-order representation. With those limits in view, the prediction is straightforward. If the model were responding to the specific fact that its activations had been altered, the gaslight condition should produce fewer injection reports than genuine steering.

It did not. Given the two original response options, Llama-3.1-70B reported an injection on 23.5 percent of control trials, 80.5 percent of steered trials, and 87.5 percent of gaslight trials. Qwen3-32B reported an injection on 18.5 percent of control trials, 85.1 percent of steered trials, and 92.9 percent of gaslight trials. Gemma-3-27B reproduced the original two-way pattern to a large extent and failed the three-way test, while Qwen-2.5-72B almost never reported an intervention in any condition and did not reproduce the original effect. When prompt manipulation was added as a third option, Llama-3.1-70B correctly identified gaslight trials as such on 27.5 percent of them and chose the activation-intervention option on 41.6 percent, and it assigned 40.8 percent of genuinely steered trials to the prompt-manipulation option. Qwen3-32B chose the activation-intervention option on 47.5 percent of gaslight trials and the correct option on 22.6 percent (Singh et al., 2026). The authors also report that in the three-way

setting some of the apparent detection is a preference for a particular option regardless of the intervention type. This pattern is consistent with an anomaly reading, since the reports did not discriminate the source of the perturbation, but it is not diagnostic of that reading, since instruction following, response-option bias, mixed mechanisms, and monitoring that is indifferent to provenance remain compatible with it.

One description of this result must be corrected, because it is natural and wrong. The gaslight condition does not leave the model's internal state unchanged. A changed prompt changes the subsequent computation, and the internal state during gaslight trials differs from the state during neutral trials. The authors state the point exactly: the input might cause the model's state to be unusual, but nothing is injected into its activations. The finding is that a similar report arises through a different causal route, one that differs at the input stage but ends, like steering, in an altered internal state, so that what differs is the provenance of the alteration. The report is therefore insufficient to identify activation injection specifically. Treating prompt processing as external to the model's internal dynamics would misdescribe the control and would obscure its actual force, which is that the report fails to track the variable it was supposed to track.

The control also shows why base-rate checks on unmodified prompts do not settle the anomaly question. A model that rarely reports an injection on neutral control trials, which Macar and colleagues report as a zero false-positive rate in their setting, has shown only that its reports are not produced indiscriminately, which is a behavioral result and carries no commitment about mechanism. It has not shown that the reports are specific to activation-level perturbation, because the relevant alternative is an input-level perturbation that is atypical in the same way. The right comparison class for a false positive is a matched manipulation that produces an atypical state by a different route.

Three cautions apply to the interpretation of the gaslight result. The experiments used open-weight models, not the Claude models studied in the original work, whose internals are not accessible outside Anthropic, so the result does not directly refute anything about Claude and bears on the paradigm only to the extent that the relevant behavior is shared. The authors themselves conclude not that the models demonstrably lack introspective capacities, since models could likely be trained to separate the two interventions, but that success in the two-way paradigm cannot on its own be taken as evidence of such capacities. And a press summary of the work states that the models were merely reacting to anything unusual, which is a stronger statement than the data support. The data show that an alternative explanation is available and consistent with the results. They do not show that it is the operative one. The authors add a further limit: their own experiments test a strictly weaker condition, whether reports track interventions on hidden states specifically rather than irregularity in general, and a model that passed their three-way test would still not thereby be shown to deploy meta-representations.

5 Label Prediction and Privileged Access

The second family of studies raises a different issue. In these studies a proxy is derived from a model's hidden states, either from a probe trained on labeled data or from a principal component, and the model is then asked to predict the resulting labels for new inputs in context. Ji-An and colleagues (2025) took successful prediction as evidence of introspection, using first-person descriptions of actions labeled for ethical violations. Singh and colleagues asked whether the model could succeed by picking up on the task implied by the labels. When the labels were randomly permuted before the probe was trained, which preserves the label distribution and still defines a well-formed direction in hidden-state space, accuracy fell to near the majority-class baseline. For the principal-component labels, probes trained on the layer-0 representations of the inputs alone tracked the model's in-context performance, so the labels were predictable from input features without access to the contextualized hidden states from which the target labels were derived. This establishes a possible shortcut and not a demonstrated one, since decodability of a label from the input does not show that the model uses input features to produce its prediction. The conclusion that follows is the narrower one that the original success does not require privileged access. The authors add that the related claim that models can control their neural proxy is equally consistent with the ability to control generation, which models are known to have.

For the belief-dominance task of Steinmetz Yalon and colleagues (2026), in which prompts such as one asserting that the capital of France is Rome are scored by the relative internal strength of the known and counterfactual answers, linear probes given only the uncontextualized embeddings of the two entities matched or exceeded the models' in-context accuracy on Llama-3-70B and Gemma-3-27B. The authors suggest that simple properties of the entities, such as their frequency in training data, may account for the success. They are also explicit that the result shows only that the reported findings do not require privileged access, and not that no version of these paradigms could satisfy that condition.

The two controls are of different kinds. Random relabeling is a check on construct validity, asking whether the task can be solved by a semantic route that does not require access to the particular hidden-state property being tested. The layer-0 probe is a check on information locality, asking whether the label is recoverable from features of the input. Both generalize beyond this literature to any task in which an experimenter derives labels from a model's internals. The failure they reveal is related to the first but distinct from it. The label-prediction studies raise the question whether privileged access is needed to solve the task at all. If a probe given only the input does as well as the model, the task does not require information that the input does not contain, and the result carries no implication about the model's access to its own states. The injection studies, by contrast, supply evidence of privileged access when the intervention varies independently of the prompt, since the injected vector is not in the input,

and the question they raise concerns what kind of processing that access supports. Combining the two under a single heading would flatten a distinction that organizes the whole problem. One failure concerns whether the report can depend on hidden information, and the other concerns what the dependence consists of.

6 Evidence That Discriminates Among Explanations

If the standard of Section 3 is accepted, the constructive question is which designs would make the readings predict different outcomes. Five kinds of evidence deserve attention, and they differ in weight. The first two are diagnostics for the anomaly and direct-effect readings, the third is a control on privileged access, the fourth is the substantive criterion for second-order representation, and the fifth is the evidential standard that ties the others together.

The first is matched anomaly across routes. Perturbations delivered by activation steering, by prompt content, by corrupted context, and by distribution shift in the input can be matched on some measure of atypicality, for instance the model's own predictive surprise, though token-level surprise is an imperfect matching measure because it need not capture the activation-level atypicality that a detector responds to. If reports track the route and not the atypicality, the anomaly reading loses force, although a first-order detector could also be tracking a distinctive steering signature. If they track atypicality and not the route, as the gaslight result is compatible with, the interpretation of the reports as specific to activation intervention loses support, although a monitor of abnormal processing could also respond to abnormality whatever its source. Each outcome therefore leaves a residual alternative, and the matched-anomaly design becomes substantially more informative when combined with the downstream-use evidence described under the fourth kind below. Alone it narrows the interpretation of the reports without settling the mechanism. Because source specificity is orthogonal to second-order status, the matched-anomaly diagnostic bears on provenance-representing monitors specifically and not on monitors in general.

The second concerns the direct-effect reading, and here the honest statement is that a clean dissociation design is not currently obvious. That reading predicts that reports depend on how strongly the intervention promotes concept-related output. Neither holding the target representation fixed nor varying it is sufficient alone. A useful dissociation would manipulate the candidate reporting pathway while preserving the target condition, and separately vary the target condition while preserving ordinary task performance, testing whether the resulting pattern supports a monitoring role. Such a design presupposes that the candidate pathway can be identified independently of the report, and that identification is itself part of what is in question. The evidence-carrier and gate features reported by Macar and colleagues are candidate identifications of such a pathway for the detection report, but they were found through their role in producing that report, so treating them as the independent pathway would

risk circularity unless their role were established by interventions that do not route through the reporting task. This is an open design problem and not a solved one.

The third is the relabeled or shuffled control already used by Singh and colleagues. Random relabeling, which preserves the training-label counts and still defines a well-formed if arbitrary direction in hidden-state space, disrupts the original semantic correspondence between inputs and labels, testing whether the reported success survives when that correspondence no longer supplies an obvious shortcut. It does not establish that all input-based prediction is impossible, since a trained probe's output remains a function of the activations that the input elicits. Stronger evidence of privileged access would come from varying a target internal condition independently of the visible input and testing whether reports track that variation. Any comparison with an input-only baseline should also specify the baseline's access restrictions and resource budget, and informational exclusion should be kept distinct from computational difficulty. An independently randomized, unannounced intervention introduces trial-specific information absent from the prompt, whereas predicting a deterministic hidden-state property may only require more computation than the baseline can perform, which shows an advantage under restrictions without establishing privileged access in the same sense. For a deterministic model without intervention, the activations are fully determined by the input, so the relevant contrast concerns the information and computation available to the comparison system and not absolute recoverability.

The fourth concerns what would make a state about the model's own processing. Singh and colleagues use an analogy to reflection in programming languages: an ordinary program uses values in memory, whereas a reflective program can also obtain descriptions of those values and act on them. A rule that prints that the activations were altered whenever a hidden-state norm exceeds a threshold could reliably detect interventions through a direct causal pathway, without monitoring the state as such, and would produce behavior identical to a genuine monitor. They draw on Ramsey's (2007) job description challenge: calling a state a representation should explain something that is not already explained by calling it a reliable causal relay, as a bimetallic strip covaries with temperature but is a representation of temperature only if the system uses it as a stand-in. They list properties that would provide stronger, though still defeasible, evidence. The feature should generalize in a manner appropriate to its proposed content, so that a representation of abnormal processing generalizes across relevant kinds of abnormality. Downstream processes should use the feature as information about the model's processing, for calibration, hedging, or control. And the proposed content should be distinguishable from the cue that activates the feature, which for a norm detector means testing cases of high norm without an intervention and intervention without high norm. These properties can also be exhibited by a first-order anomaly signal used for calibration or control, so a further requirement is needed: downstream computation must treat the signal as information about the reliability, provenance, or condition of another computational process, and not merely

respond to it. Whether such treatment occurs remains an empirical and theoretical task. These criteria apply to representation generally. Successful probing establishes decodability, while causal interventions and downstream use can provide additional evidence that the decoded information plays the proposed representational role, and the same demand is made of any putative representation. The requirement is therefore a matter of degree, and the interesting question is how much downstream integration, generalization, and cue-independence suffice. The standard is demanding but not impossibly so, because the same evidence is routinely gathered for other claims about representation in these models. This is an operational reading of the second-order requirement, and it is the one the standard of this essay adopts. Without a reading of this kind the standard would replace an ambiguous behavioral criterion with an equally ambiguous representational one.

The fifth is convergence of behavioral and mechanistic evidence. Singh and colleagues state that task performance alone cannot satisfy the second-order condition, which requires designs under which first-order and second-order accounts make divergent predictions and, ultimately, converging evidence about the model's internal organization. They observe that human metacognition research must infer mechanism from behavioral dissociations, whereas the internal computations of a language model are available for direct recording and intervention in accessible models, which makes converging evidence a realistic standard. The standard proposed here is compatible with that view. To the extent that behavior can be arranged so that rival mechanisms make divergent predictions, behavior bears on the functional question, and the mechanistic evidence then addresses what the behavioral dissociation cannot.

7 Limits of the Inference

Two further limits concern the systems tested and the scope of the original claims. The first concerns models and training. The replications in open-weight models and the original experiments in Claude models are different systems, and capabilities reported to depend on post-training (Macar et al., 2026) may differ across them. Singh and colleagues also restrict attention to pretrained and instruction-tuned models, because finetuning for a task may install a task-specific mechanism and not reveal a general capacity, so results from models finetuned explicitly for steering awareness address a different question. Conclusions should be indexed to the models tested.

The second concerns the original claims, which were more qualified than their popular summaries. Lindsey's paper describes the observed abilities as limited and context-dependent, concedes that the mechanisms involved may be shallow or narrowly specialized, and disclaims conclusions about subjective experience (Lindsey, 2025). It also reports experiments beyond injection detection, including recall of prior internal representations as distinct from text inputs and the modulation of internal representations under instruction. A reexamination of one

paradigm does not dispose of the others, and each needs its own control analysis. A further limit is theoretical: Singh and colleagues target the family of views on which a system knows its states through a dedicated capacity to inspect them. On inferential or transparency views of self-knowledge, which they note are easier to satisfy, the question has a different shape, and nothing in this essay decides between the families.

8 A Standard for Self-Report Claims

The argument can be generalized without a survey of every claim of machine self-report. A verbal report about an internal state can be asked to support three inferences, and these differ in kind as well as in strength. The first is that the report depends causally on the state, which intervention experiments establish and which is compatible with every reading in Section 2. The second is privileged access, that the dependence runs through information absent from the input, as when an independently randomized, unannounced intervention introduces trial-specific information, as distinct from an advantage that reflects only computation a restricted comparison system could not afford. The injection paradigm supplies evidence of privileged access when an unannounced intervention varies independently of the prompt, and the label-prediction results, as deployed, do not establish it. The third is second-order representation, that the report is produced by a state whose content concerns another state or process of the model, which requires the kinds of evidence described in Section 6. A fourth property, specificity to the source of the perturbation, is a diagnostic for the anomaly reading and is orthogonal to these three, as argued in Section 4.

Two consequences follow. First, the three questions are distinct dimensions of assessment and not a necessary ladder, since a second-order representation could concern a deterministic state and so fail strict informational exclusion, while a direct steering effect could satisfy that criterion without any monitoring. None of these findings, considered alone, establishes the others. Second, even convincing evidence of functional self-monitoring would require a separate argument before it supported claims about consciousness or welfare, since a monitoring process of the kind described is a claim about the organization of computation, and the step from that organization to experience is not an empirical consequence of the experiments discussed here.

The practical value of stating the standard this way is that it directs effort. Disputes over whether models introspect tend to proceed by multiplying demonstrations or multiplying objections. If the problem is identification, the productive work is the design of experiments and the collection of mechanistic evidence in which the rival mechanisms predict different results, and the reports of both sides can then be evaluated against what that work is able to show.

A Identification and the Orthogonality of Source Specificity

This appendix states the identification problem of Sections 2 to 4 in a form that can be checked, and its scope should be stated at the outset. It formalizes report-level distinguishability under stipulated mechanism models. Distinct internal architectures can implement exactly the same report function over every reachable state, so distinguishing causal organizations requires interventions on intermediate variables, which a report-level analysis cannot supply. Let X be a prompt drawn from a distribution P_X, and let $I \in \{0, 1\}$ record whether the target activation intervention was applied, assigned by the experimenter independently of X. Let $G \in \{0, 1\}$ record whether the prompt carries a prompt-level manipulation, so that G is a function of X. For fixed model parameters the hidden state is a deterministic function $H = F(X, I)$, and the report $R \in \{0, 1\}$, with $R = 1$ meaning that an injected thought is reported, is sampled with probability $P(R = 1 \mid H = h) = \rho(h)$. A mechanism model is a hypothesis about the form of ρ. The direct-effect, anomaly, and monitoring readings of Section 2 are respectively $\rho_D = \sigma_D \circ d$, $\rho_A = \sigma_A \circ a$, and $\rho_M = \sigma_M \circ m$, where d, a, and m are summaries of the state (strength of concept promotion, atypicality, and the content of a candidate monitoring representation) and the σ are link functions. Such a factorization distinguishes causal organizations only to the extent that the summaries and links carry independent restrictions. A condition δ of a design specifies a distribution over (X, I), with state support $S_\delta = \operatorname{supp} F(X, I)$, and a design $\mathcal{D}$ is a set of conditions with $S_\mathcal{D} = \bigcup_{\delta \in \mathcal{D}} S_\delta$.

Proposition 1 (Separation requires distinguishing states)**.** *If two mechanism models k and l satisfy $\rho_k(h) = \rho_l(h)$ for every $h \in S_\mathcal{D}$, then they predict the same report distribution under every condition of $\mathcal{D}$. If $\rho_k - \rho_l$ has constant sign on $S_{\delta'}$ and δ' assigns positive probability to states where it is nonzero, then they predict different report distributions under δ'.*

Proof. Under a condition δ, $P_\delta(R = 1) = \mathbb{E}_\delta[\rho(H)]$. If $\rho_k = \rho_l$ on S_δ the expectations coincide. If $\rho_k - \rho_l$ has constant sign on $S_{\delta'}$ and is nonzero with positive δ'-probability, then $\mathbb{E}_{\delta'}[\rho_k - \rho_l] \neq 0$. □

The proposition preserves a distinction that the exposition should not blur. Pointwise disagreement between ρ_k and ρ_l is necessary for separation but not sufficient for it, because mechanisms that disagree at individual states can have identical aggregate report rates through cancellation, as Proposition 2 shows, and the constant-sign condition is a sufficient condition that excludes cancellation. In the original two-way design $\mathcal{D}_2 = \{\delta_0, \delta_1\}$, with δ_0 the unmodified forward pass and δ_1 the steered one, intervention status and atypicality are not independently varied, so the anomaly summary a and a monitoring summary m may agree in the sense that $\sigma_A \circ a = \sigma_M \circ m$ on $S_{\mathcal{D}_2}$ while differing elsewhere. The gaslight condition δ_G changes the distribution of inputs and of the resulting states. Whether it adds states outside $S_{\mathcal{D}_2}$ is an additional assumption, since different prompts can in principle map to overlapping states, but

it is on states where a and a provenance-sensitive summary differ that the added condition can separate them, and the empirical question is whether such states are reached.

Proposition 2 (Aggregate equality does not imply pointwise equality)**.** *There exist mechanism models k and l and a condition δ with $\rho_k \neq \rho_l$ on S_δ such that $P_\delta(R = 1)$ is the same under both.*

Proof. Let $S_\delta = \{h_1, h_2\}$ with each state having probability $1/2$ under δ, let $\rho_l \equiv q$ for some $q \in (0, 1)$, and let $\rho_k(h_1) = q + \varepsilon$ and $\rho_k(h_2) = q - \varepsilon$ with $0 < \varepsilon < \min\{q, 1 - q\}$. Then $\mathbb{E}_\delta[\rho_k] = q = \mathbb{E}_\delta[\rho_l]$ although the functions differ at both states. □

Corollary 1 (Equivalent achievable distributions)**.** *Let $\mathcal{P}_k(\mathcal{D})$ and $\mathcal{P}_l(\mathcal{D})$ be the sets of joint distributions of the design's observable statistics that mechanism classes M_k and M_l can generate under $\mathcal{D}$. If $\mathcal{P}_k(\mathcal{D}) = \mathcal{P}_l(\mathcal{D})$, then no decision rule based on those observations alone is uniformly reliable in distinguishing the two classes.*

Proof. Let Γ be any decision rule that maps data from $\mathcal{D}$ to a verdict k or l, and let P be a distribution in the common set. The law of Γ's output under P is the same whichever class generated P. A rule that returns k with high probability under P is reliable when M_k is true and unreliable when M_l is true, and the reverse holds if it returns l. Hence no rule is reliable for both classes at P. □

Remark 1. The corollary concerns what a design can separate, which is a frequentist question of achievable distributions. It does not claim that Bayesian comparisons are uninformative: two hypotheses with the same achievable distributions can assign them different prior weights, and the data can then favor one through its prior predictive distribution. Identification in the sense of this essay is the weaker requirement that the achievable distributions differ, and not a claim about posterior weights. Whether a class with a threshold and a slope in its link function can fit a given pair of report rates depends on the summary values, the link, and any parameter restrictions, and the observable summary of a two-way experiment is only a pair $(P(R = 1 \mid \delta_0), P(R = 1 \mid \delta_1))$. The availability of a fit does not by itself discriminate mechanism classes.

Observation 1 (Conceptual independence of source specificity and second-order status)*.* Let a policy be *source specific* if there is a level u of atypicality, at which states arising by both activation intervention and prompt content occur, such that $P(R = 1 \mid \Pi = \text{steer}, a(H) = u) \neq P(R = 1 \mid \Pi = \text{prompt}, a(H) = u)$, where Π records how the state arose. Let a policy be *second order* if its report is produced through a state whose content concerns another state or process of the same system, as determined by criteria that are justified independently of the report, such as those of Section 6. Neither property entails the other provided that two possibility claims hold: first, that there can be a monitor satisfying those criteria whose reports

are not source specific, and second, that there can be a source-specific detector that fails them. A criterion's silence about provenance does not by itself establish either claim, since a property can entail another without mentioning it in its definition.

Argument. This is a conditional conceptual claim, and the two possibility claims are its premises. The argument supports them with examples, but their force depends entirely on the independently justified criteria doing the work, and it is not a construction that settles representational status. Suppose a monitor is independently justified as second order, by downstream use, generalization, and cue-independence. Nothing in the definitions requires its content to include provenance, so it may represent that the current state is atypical without representing how the atypicality arose, and its reports are then not source specific. Conversely, a stipulated first-order detector of a distinctive steering signature, such as a feature of the added vector's direction, produces reports that differ by provenance at matched atypicality without any state whose content concerns another process, and its reports are then source specific. Source specificity therefore does not suffice for second-order status, and its absence does not rule that status out. The claim is the absence of an entailment, and whether any particular system falls in one cell or another remains an empirical question. □

B Privileged Access as Informational Exclusion and as Resource-Relative Advantage

This appendix separates two senses of privileged access that Section 6 distinguishes. Write H for Shannon entropy, reserving H for hidden states. Let Z be a variable describing the target internal condition, X the visible input, and R the report. Informational privileged access of R about Z is the condition $I(R; Z \mid X) > 0$, where $I(\cdot;\cdot \mid \cdot)$ is conditional mutual information.

Proposition 3 (Deterministic states carry no informational advantage)**.** *Suppose the hidden state is $H = F(X)$ for a deterministic F, with no intervention, and the report is generated from (H, ξ) with ξ independent noise. Then for any property $Y = g(H)$, $I(R; Y \mid X) = 0$.*

Proof. Given X, the variable Y is constant, so $\mathsf{H}(Y \mid X) = 0$ and therefore $I(R; Y \mid X) = \mathsf{H}(Y \mid X) - \mathsf{H}(Y \mid R, X) = 0$. □

Proposition 3 is easy, and its role is to motivate the budget framing and not to establish anything about particular models. Any advantage of a model over an input-restricted comparison system in predicting a deterministic hidden-state property is therefore computational and not informational. It can be measured relative to a resource budget. For a budget B, let $\mathrm{acc}_B(Y \mid X)$ be the best accuracy in predicting Y attainable by any predictor with access to X alone and cost at most B, and let $\mathrm{Adv}_B = \mathrm{acc}(R \to Y) - \mathrm{acc}_B(Y \mid X)$ be the advantage of the model's report.

Proposition 4 (Budget-relative advantage is nonincreasing in the budget)**.** Adv_B *is nonincreasing in* B*. Suppose* $Y = g(F(X))$ *is deterministic and the comparison system is permitted to compute* F *and* g*, which deterministic dependence alone does not grant. Then a predictor computing* $g \circ F$ *on* X *attains accuracy one, so the advantage of the report is at most zero for any budget at which that computation is affordable.*

Proof. Enlarging the budget enlarges the class of admissible predictors, so acc_B is nondecreasing in B and Adv_B is nonincreasing. The second claim holds because the predictor computing $g \circ F$ is exactly correct and is admissible whenever its cost does not exceed B. □

The proposition does not claim that a single finite budget suffices for arbitrary input lengths. An unrestricted predictor with access to the relevant computation achieves perfect prediction, and whether a finite budget suffices depends on the cost of computing $g \circ F$. The contrast with an unannounced, randomized intervention is sharp. Suppose $Z \in \{0, 1\}$ is assigned independently of X with $0 < P(Z = 1) < 1$, and $H = F(X, Z)$.

Proposition 5 (Randomized interventions yield informational exclusion at every budget)**.** *For any predictor* $\hat{Z} = \phi(X)$*, regardless of cost,* $P(\hat{Z} = Z) \leq \max\{P(Z = 0), P(Z = 1)\}$*. If the report satisfies* $P(R \mid X, Z = 1) \neq P(R \mid X, Z = 0)$ *on a set of positive probability, then* $I(R; Z \mid X) > 0$*.*

Proof. Independence of Z and X gives $P(\phi(X) = Z) = \sum_z P(Z = z)P(\phi(X) = z) \leq \max_z P(Z = z)$. The second claim follows because $I(R; Z \mid X) = 0$ if and only if R and Z are conditionally independent given X. □

Remark 2. Informational exclusion does not establish second-order representation. A direct steering detector also satisfies $I(R; Z \mid X) > 0$, so Proposition 5 certifies that the report depends on information absent from the input and says nothing about how that information is represented, which is the content of Observation 1.

Propositions 3 to 5 formalize the distinction drawn in the text: evidence of privileged access in the strict sense requires a source of trial-specific information that the input lacks, whereas an advantage over a restricted comparison system on a deterministic quantity is an advantage under a budget and shrinks as the budget grows.

C The Relabeling and Input-Probe Controls

Let the data be $(x_i, y_i)_{i \leq n}$ with semantic labels y_i, let $h(x)$ be the hidden state at a layer, and let a probe with weight w be trained to predict y. The proxy labels used in self-report classification are $z_i = \mathbf{1}[\langle w, h(x_i) \rangle > \tau]$, or cluster assignments derived in the same way. In the relabeling control, y is replaced by $y_{\pi(i)}$ for a uniformly random permutation π, a probe w_π

is trained on the permuted labels, and proxy labels z_i^π are derived from it. The permutation preserves the training-label counts, and $\langle w_\pi, h(x)\rangle$ is a well-defined direction in hidden-state space, so z^π is a legitimate target for privileged access.

Proposition 6 (Semantic predictors are at baseline under independence)**.** *Suppose that on held-out inputs Z^π is independent of the semantic label Y. Then any predictor $\hat{Z} = \varphi(Y)$ that uses the input only through Y has accuracy at most* $\max_z P(Z^\pi = z)$.

Proof. Independence gives $P(\varphi(Y) = Z^\pi) = \sum_z P(Z^\pi = z)P(\varphi(Y) = z) \leq \max_z P(Z^\pi = z)$. □

The independence premise is a limitation of the control itself, and not only of inferences drawn from it. It must be verified empirically on held-out data, so the relabeling control functions as a diagnostic whose premise can fail, not as a test that guarantees a semantic route is excluded. The premise should be measured and not inferred from the permutation procedure. A probe trained on permuted labels can still produce held-out predictions correlated with the original semantic classes through finite-sample effects, the geometry of the representation, or regularization, and permutation preserves the training-label counts without guaranteeing either independence from the semantic label or the distribution of the resulting proxy labels. The proposition also concerns predictors that use the input only through the semantic label. It does not exclude other input-based predictors, which is why the control is a check on construct validity, asking whether a semantic route accounts for success, and not a proof that no input-based route exists.

The second control is a check on information locality. Let $\hat{z}_0$ be a linear probe trained on layer-0 representations $h_0(x)$, the input embeddings, to predict z, and let $\mathrm{acc}_{\mathrm{ICL}}$ and acc_0 denote the accuracy of the model's in-context prediction and of $\hat{z}_0$ on held-out inputs.

Observation 2 (What a matching layer-0 probe shows)*.* If $\mathrm{acc}_{\mathrm{ICL}} \leq \mathrm{acc}_0 + \epsilon$, then the model's accuracy does not exceed, by more than ϵ, what a linear function of the input embeddings attains, and the data do not establish that the model's prediction draws on information beyond those embeddings.

This establishes a possible shortcut and not a demonstrated one. Decodability of z from h_0 is a statement about information, whereas use is a statement about causal dependence. What would establish use is a causal intervention, as in the following definition.

Definition 1 (Use)**.** The model *uses* a feature $\phi(x)$ in predicting z if intervening on ϕ while holding the remaining features of the input fixed changes the prediction, that is, $P(\hat{Z} \mid \mathrm{do}(\phi = \phi'), \mathrm{rest}) \neq P(\hat{Z} \mid \phi, \mathrm{rest})$ for some ϕ'.

A matching layer-0 probe shows that the shortcut is available, so the original result does not require privileged access, and it does not show that the shortcut is the route the model takes. The conclusion drawn in the text is accordingly the narrower one.

D A Factorial Design for Cue-Independence

Section 6 requires that the content of a candidate representation be distinguishable from the cue that activates it. Let $f(h)$ be a scalar feature, for instance the activation of a candidate detector, let $c(h)$ be a scalar candidate cue such as the norm of a hidden state at a given layer, and let $I \in \{0, 1\}$ record the presence of the target steering operation. Distinguish the scalar c from any binned category $C = b(c)$. Define the intervention contribution at an exact cue value k, assuming the conditional distributions are defined and the two arms have common support at k, by

$$\Delta_I(k) = \mathbb{E}[f \mid I = 1, c = k] - \mathbb{E}[f \mid I = 0, c = k],$$

and the cue contribution at intervention level i by $\Delta_c(i; k, k') = \mathbb{E}[f \mid I = i, c = k'] - \mathbb{E}[f \mid I = i, c = k]$.

Proposition 7 (A pure cue detector has no intervention contribution)**.** *If $f = \varphi(c)$ for some function φ, then $\Delta_I(k) = 0$ at every cue value k in the common support, and if the distribution of c is the same under $I = 1$ and $I = 0$, then $\mathbb{E}[f \mid I = 1] = \mathbb{E}[f \mid I = 0]$. Hence a nonzero $\Delta_I(k)$ at a matched cue value, or a nonzero difference in means under matched cue distributions, falsifies the hypothesis that f is a function of the cue alone.*

Proof. Conditional on $c = k$, $f = \varphi(k)$ is fixed, so its conditional mean does not depend on I. If c has the same distribution under both arms, then $\mathbb{E}[\varphi(c) \mid I = 1] = \mathbb{E}[\varphi(c) \mid I = 0]$. □

The proposition does not extend to coarse categories. If $f = \varphi(c)$ and the comparison conditions only on a bin $C = b(c)$, intervened states may have values of c concentrated near one end of the bin and control states near the other, so that a pure cue detector yields a nonzero difference within the bin. Matching must therefore be at exact values or of the whole distribution of c within each comparison.

The converse of the proposition does not hold: $\Delta_I(k) \neq 0$ shows that f carries information about the intervention beyond the cue, but it does not show that f is a second-order representation, for the reason given in Observation 1. A first-order detector of a steering signature also has $\Delta_I(k) \neq 0$. Cue-independence is therefore a necessary check for the claim that a feature represents an intervention and not a correlated cue, and it must be combined with the other criteria of Section 6.

The implementation needs care. Rescaling a neutral hidden state is itself an activation intervention, so it cannot furnish an $I = 0$ condition if I meant any activation intervention, which is why I is defined here as the presence of the target steering operation and the cue manipulation is treated as a separate factor. A cue manipulation, such as rescaling a neutral state or adding a norm-matched random vector, may change other properties of the state and must be assessed for those effects. A low-strength steering vector likewise does not automatically

provide a useful condition with a low cue, because it may also weaken the target condition whose detection is being tested.

Finally, let B be a set of behaviors distinct from the report, such as hedging or calibration, and let $\mathrm{do}(f = f')$ denote an intervention that sets the feature to a value f'. A downstream-use criterion requires that $P(B \mid \mathrm{do}(f = f'))$ vary with f', and that the variation be of the form appropriate to the condition the feature is claimed to represent, so that, for instance, raising f reduces the model's expressed confidence in outputs generated under the manipulated condition. Whether a given effect of this kind constitutes treatment of the signal as information about the reliability or provenance of another process, and not a mere response to it, remains an empirical and theoretical question, as the body of the essay states.

References

Ji-An, L., Xiong, H.-D., Wilson, R., Mattar, M. G., and Benna, M. K. 2025. Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations. arXiv preprint arXiv:2505.13763.

Lederman, H., and Mahowald, K. 2026. Emergent Introspection in AI Is Content-Agnostic. arXiv preprint arXiv:2603.05414.

Lindsey, J. 2025. Emergent Introspective Awareness in Large Language Models. Anthropic, published October 29, 2025; posted to arXiv in January 2026 as arXiv:2601.01828.

Macar, U., Yang, L., Wang, Z., Wallich, P., Ameisen, E., and Lindsey, J. 2026. Mechanisms of Introspective Awareness. In Proceedings of the 43rd International Conference on Machine Learning, PMLR 306:85403–85429.

Ramsey, W. M. 2007. Representation Reconsidered. Cambridge University Press.

Singh, S., Linzen, T., and Ravfogel, S. 2026. Can LLMs Introspect? A Reality Check. In Conference on Language Modeling (COLM 2026); arXiv preprint arXiv:2605.26242v2.

Steinmetz Yalon, N., Goldstein, A., Mudrik, L., and Geva, M. 2026. Indications of Belief-Guided Agency and Meta-Cognitive Monitoring in Large Language Models. arXiv preprint arXiv:2602.02467.