Evidentiary Authority and the Acceleration of AI: Safety Claims, Attribution, and Revisable Judgment

Flyxion

Independent Researcher

November 2026

Abstract

Two practices share a structure. Detectors of machine-generated content classify an artifact by its apparent production mechanism and then treat that classification as a verdict on authorship, originality, or truth. Commercial developers of advanced AI systems describe themselves as responsible and then treat that description as a verdict on the adequacy of their safeguards. In both cases a property of the source is substituted for an assessment of the claim, its evidence, and the standing of the party making it. This essay argues that the substitution is a recurring failure of institutional judgment, with a consequence in the second case that the first lacks: authority acquired through a safety reputation can be converted into capital, and capital into faster capability development. The argument has four parts. First, a non-identifiability result shows that artifact-only classification cannot recover intellectual provenance when distinct production histories yield the same artifact. Second, a claim graph with explicit evidence, warrants, objections, and dependencies is proposed as a source-neutral alternative, with a soundness condition for revision when premises are defeated. Third, a documented record of one commercial case is assembled from public filings and reporting, separating what it establishes from what it leaves open. Fourth, a comparative-statics model of safety-branded competition identifies the conditions under which safety reputation lowers or raises aggregate risk, and shows that even genuine safeguards carry an induced acceleration term that must be weighed against their direct effect. It is extended to a provider's position as repository of private intellectual work and supplier to governments, distinguishing capacity to observe from its exercise. No claim is made about the sincerity of any researcher or the value of safety work as such. The contribution is a method for stating, testing, and revising such claims, with a protocol by which it can fail.

Keywords: provenance, defeasible reasoning, claim graphs, truth maintenance, AI-generated content detection, safety claims, institutional incentives, epistemic infrastructure

1 Introduction

A reader who meets a competent essay and is told that a detector assigns it a high probability of machine generation has learned something about the statistical texture of the text and almost nothing about who conceived its argument, whether its premises are true, or whether its author deceived anyone. The detector has measured a feature of the artifact. The institution that acts on the score, whether a journal, a platform, or a school, usually does so as though a feature of the artifact were a feature of the history that produced it, and as though the history settled the questions that matter. The same slippage occurs in a more consequential form when a company that builds powerful systems is described, and describes itself, as the safe alternative. A property of the source, here its reputation for caution, is carried over to a claim about the world, here that its systems are adequately constrained. In both cases the unit of evaluation has been quietly changed from the claim to the claimant.

This essay treats the two cases together because they share a remedy and because the second inherits a difficulty from the first. The remedy is to make the claim the unit of adjudication, to attach to it the evidence and warrants on which it rests, to record the objections that have been raised against it, and to require that its status change when the support beneath it changes. The difficulty is that a source-neutral method cannot ignore sources altogether. Testimony, measurement, and provenance records are themselves evidence, and their reliability depends on where they came from. The method developed here therefore separates four things that ordinary usage runs together: the validity of a proposition, the credibility of a source, the authorship of a text, and the authorization of a party to speak for an institution or to take credit for a result.

The second case adds something the first lacks. A misclassified essay harms its author and misleads its audience, but it does not alter the pace at which essays are written. A safety claim made by a firm that is also raising capital for capability development can alter that pace, because the reputation conferred by the claim is an input to the firm's access to financing. Whether this effect is large, small, or outweighed by the real safeguards the same reputation supports is an empirical question. The argument of this essay is that the question has a determinate structure, that the structure can be written down, and that the institutions now asking the public to rely on safety claims have not yet written it down.

The sections that follow proceed in order. Sections 2 through 4 develop the non-identifiability of provenance, the semantic dependence of misrepresentation, and the

principle of source neutrality with its limits. Section 5 introduces the claim graph and states conditions for sound revision. Section 6 treats attribution and authority as institutional objects. Section 7 assembles the public record for one commercial case. Section 8 presents the comparative-statics model. Section 9 applies the claim graph to safety claims directly. Section 10 turns to anthropomorphic mediation and the concentration of epistemic infrastructure, and Section 11 to the pressure that states can bring to bear on a privately governed provider. Section 12 sets out an evaluation protocol, and Section 13 states limitations.

2 Non-Identifiability of Provenance from Surface Features

Let H be a set of production histories, each recording who contributed which ideas, which tools transformed which drafts, and in what order. Let A be the set of artifacts, and let $F : H \to A$ map a history to the artifact it yields. Let $O : H \to \Omega$ extract a specified intellectual-origin property, for example whether the central argument was conceived by a particular person before any generative system was consulted. An artifact-only estimator is a function $\hat{O} : A \to \Omega$.

Proposition 1 (Non-identifiability of intellectual provenance)**.** *If there exist histories $h_1, h_2 \in H$ with $F(h_1) = F(h_2)$ and $O(h_1) \neq O(h_2)$, then no artifact-only estimator $\hat{O}$ satisfies $\hat{O}(F(h)) = O(h)$ for every $h \in H$. If histories are drawn from a distribution, the error probability of any such estimator on an artifact a is at least one minus the largest posterior probability that any single origin value has given a, a quantity fixed by the population of histories and not by the quality of the estimator.*

Proof. Since $F(h_1) = F(h_2) = a$, the estimator returns one value $\hat{O}(a)$. That value differs from at least one of $O(h_1)$ and $O(h_2)$. For the distributional statement, the best response to a is the posterior mode, and every other response errs with at least the complementary probability. The posterior is a property of how histories map to artifacts in the population, and given that population no refinement of the estimator can improve on it. A better model of the population, with more recorded histories and a better prior, does improve it, and that is what provenance records are for. □

The result is elementary, and its force lies in the empirical premise that the required pairs of histories exist. They do. A person who drafts an argument by hand and then asks

a system to improve its prose may yield an artifact indistinguishable, in every respect a detector can read, from one in which the system supplied the argument and the person supplied only a request. A person who dictates, a person who translates, and a person who has the text reformatted by a program all produce artifacts whose surface depends on a later transformation and not on the earlier act of conception. Empirical work on detector reliability points the same way: published analyses show that paraphrasing defeats several detectors and that some detectors misclassify the writing of non-native speakers at elevated rates [20, 12]. These findings concern the detectors' error rates, whereas Proposition 1 concerns a limit that holds even for a perfect detector of surface features. The premise of the proposition should be stated with care. Constructing distinct histories that yield an identical artifact is straightforward, and the proposition claims no more than that. Showing that a particular deployed detector cannot separate two particular real histories is a separate empirical question that the proposition does not answer.

Two consequences follow for interpretation. A detector's reported percentage is not a proportion of intellectual authorship, because intellectual contribution is not a quantity distributed over the sentences of a finished text. A passage rewritten by a system may carry a contribution wholly the person's, and a passage left untouched may carry little. The second consequence concerns responsibility. Textual realization, intellectual contribution, editorial transformation, and responsibility are four different relations between a party and an artifact. A person may be responsible for a text they did not type, and a system may type a text for which it is not responsible in any sense that matters to an institution. Collapsing the four into one score produces decisions that cannot be defended on any of the four grounds.

3 The Semantic Dependence of Misrepresentation

The same structure appears in the evaluation of images. A photograph that is authentic in every technical respect becomes a misrepresentation when it is attached to a caption asserting that it shows an event it does not show. A synthetic illustration that accurately communicates a fact established independently is not a misrepresentation merely because no camera was involved. Cropping, enhancement, and compositing occupy a range between these poles, and where a particular edit falls depends on what the surrounding text asserts the image to be.

The relevant unit of adjudication is therefore the artifact together with its asserted

interpretation and context. An authenticity check addresses one question about the pair, namely whether the pixels descend from a capture event. It does not address the question a reader needs answered, which is whether the claim the image is being used to support is warranted. Standards for content credentials, such as the specification of the Coalition for Content Provenance and Authenticity, record the history of an asset and are valuable for that purpose [4]. They record origin and transformation and do not adjudicate the truth of an accompanying assertion. A credentialed photograph under a false caption passes the check and misleads the viewer, and an uncredentialed drawing under an accurate caption fails it and misleads no one. This is not an objection to provenance standards. It is a statement of the question they answer, and of the need for a layer above them that answers the other one.

4 Source Neutrality and Evidentiary Relevance

The principle of source neutrality for propositions can be stated plainly. If a proposition p is independent of how an artifact came to be written, then the truth value of p does not depend on that history. A correct proof is correct whether a person, a system, or a collaboration produced it, and an incorrect proof is not rescued by a distinguished author. Institutions that decline to evaluate the proof because of the manner of its production have declined to do the work that evaluation consists of.

Source neutrality holds for propositions and does not hold for evidence. A measurement is evidence only given an account of the instrument, a report is evidence only given an account of the reporter, and a provenance record is evidence only given an account of the system that wrote it. The reliability of sources is therefore a legitimate input to adjudication, and a method that ignored it would be as defective as one that substituted it for everything else. The distinction to preserve is between using a source's reliability to weight the evidence it supplies and using a source's identity to settle the question the evidence bears on. The first is ordinary epistemic practice. The second is the substitution this essay is concerned with.

Four relations should accordingly be kept apart. Validity concerns whether a proposition follows from its premises or is supported by its evidence. Credibility concerns how much weight a source's testimony earns. Authorship concerns who conceived and who realized a text. Authorization concerns whether a party has standing to assert, endorse, or take credit for a result on behalf of itself or an institution. Each can be assessed without settling the others, and the pathologies discussed below arise when an answer to

one is offered as an answer to another.

5 Claim Graphs and Defeasible Adjudication

The proposed alternative represents judgment as a graph whose nodes are claims, evidence items, and inference steps. An inference step connects a set of premises, which jointly license a conclusion, to that conclusion, and carries a warrant stating why the premises bear on it. Representing inferences as nodes and not as edges between single claims matters, since an inference may require several premises together and a separate edge for each would wrongly suggest that each is independently sufficient. Objections are directed at claims, including evidence items whose provenance is challenged, and at inference steps, where they undercut the warrant. This architecture draws on the literature on defeasible reasoning, in which conclusions are held subject to defeat by further information [15], on abstract argumentation frameworks, in which acceptability is defined over attack relations [7], and on truth-maintenance systems, in which the dependencies of a conclusion are recorded so that its status can be recomputed when a premise changes [6, 5]. The contribution here is a particular combination of these ideas with explicit provenance and revision obligations, and the combination is offered as a method to be tested.

Definition 1 (Claim graph)**.** A claim graph is a tuple $G = (C, C_0, N, \mathrm{Prem}, \mathrm{Conc}, W, B, \pi)$ in which C is a finite set of claims and evidence items, $C_0 \subseteq C$ is the set of base items provisionally available as premises, subject to objections, N is a finite set of inference steps, $\mathrm{Prem}(n) \subseteq C$ is the nonempty premise set of a step n, $\mathrm{Conc}(n) \in C$ is its conclusion, $W(n)$ is its warrant, $B \subseteq (C \cup N) \times (C \cup N)$ is a set of objections, with $(x, y) \in B$ meaning that x objects to y, and π assigns a provenance record to each base item. Write $\mathrm{Sup}(c) = \{n \in N : \mathrm{Conc}(n) = c\}$.

Statuses are assigned by a labeling L that maps each node to one of in, out, and und. An inference step is in if all its premises are in and every objector to it is out, it is out if some premise is out or some objector to it is in, and it is und otherwise. A claim is in if it is a base item or has an in inference step, and every objector to it is out. It is out if some objector to it is in, or if it is not a base item, has at least one inference step, and every such step is out. It is und otherwise. Call these rules the operator Φ. Each rule is monotone in the information order in which und lies below both in and out, so Φ has a least fixed point, reached by iterating from the labeling that assigns und to everything. That fixed point is the grounded labeling $\Lambda(G)$, and a claim is supported if it is in under

$\Lambda(G)$, refuted if out, and unresolved if und.

Two features of this choice should be stated plainly. A cycle of mutual support with no base item among its supporters never becomes in, since the iteration never has a first step to start from, so circular justification does not manufacture support. Mutually attacking claims remain und, since the grounded labeling is the skeptical semantics. Preferred and stable semantics need not agree with it or with each other on such configurations [7], and the grounded labeling is adopted here as the default because it is unique and computable by iteration, with alternatives available as reported sensitivity analyses. The third status matters, because a method that must assign every claim to true or false will misrepresent the situations in which evidence is genuinely incomplete, which are the situations institutions most need help with.

5.1 Revision Obligations

The central requirement concerns what the adjudicator must do when the graph changes. The requirement is the dependency-sensitive form of a familiar idea from truth maintenance.

Proposition 2 (Dependency-sensitive revision)**.** *Let G' be the graph after an update and let c be a claim that is not a base item. If no inference step in* $\mathrm{Sup}(c)$ *is* in *under $\Lambda(G')$, whether because a premise is no longer* in *or because an objection to its warrant is* in*, then c is not* in *under $\Lambda(G')$, and an adjudicator that is sound must withdraw its supported status. In particular, if a premise p belongs to* $\mathrm{Prem}(n)$ *for every $n \in \mathrm{Sup}(c)$ and p is* out*, withdrawal is required. Conversely, if some step in* $\mathrm{Sup}(c)$ *is* in *and every objector to c is* out*, then c is* in *and withdrawal is unjustified.*

Proof. The labeling $\Lambda(G')$ is a fixed point of Φ. For a non-base claim, the rule for in requires an in step in $\mathrm{Sup}(c)$, and none exists, so c is not in. The special case holds because an out premise makes every step in $\mathrm{Sup}(c)$ out. For the converse, an in step together with out objectors satisfies the rule for in. □

Defeat can arrive by an undercut and not only by a premise: new evidence may defeat a warrant without touching any premise on which the warrant operates, and an objection may be directed at the conclusion itself. A system that tracks only premise status therefore violates the first half of the proposition, and the treatment of inference steps as objects of objection is designed to prevent that.

Because the first half of the proposition is nearly a restatement of the definition, the more demanding requirement concerns how a system reaches the new labeling. Recomputing the whole graph after every change is correct and may be prohibitively expensive, and an incremental update is only acceptable if it yields the same statuses. Say that a node is directly affected by an update if its own defining data differ before and after, that is, its premises, its set of supporting steps, its objectors, or its base status, and let U be the set of nodes reachable from the directly affected nodes along the dependency relation, in which a premise points to the steps it feeds, a step points to its conclusion, and an objector points to what it objects to.

Proposition 3 (Incremental correctness)**.** *Under the grounded labeling,* $\Lambda(G')(v) = \Lambda(G)(v)$ *for every node* $v \notin U$*. Moreover, the labeling obtained by recomputing only the nodes of* U*, taking the labels of nodes outside* U *from the earlier labeling, equals* $\Lambda(G')$*.*

Two structural facts about U are used in the proof. The set U is closed under successors by construction, so its complement is closed under predecessors: if $v \notin U$ had a predecessor in U, then v would be reachable from the directly affected nodes and would lie in U. It follows that any cycle that contains a node of U lies wholly inside U, since every node on the cycle is reachable from that node. A cycle can therefore never connect an unaffected node to the affected region in both directions, and an unaffected node's label never depends on the affected region.

Proof. Let L_t and L'_t be the labelings after t iterations of Φ on G and Φ' on G', both from the all-und labeling. We show by induction that $L'_t(v) = L_t(v)$ for $v \notin U$. The case $t = 0$ is immediate. For the inductive step, the new label of v is determined by its own defining data and the labels of its direct predecessors. Since $v \notin U$, its defining data are unchanged, so its predecessors in G' are its predecessors in G, and none of them lies in U. Their labels agree at stage t by hypothesis, so the label of v agrees at stage $t + 1$. Both sequences stabilize, which gives the first claim. For the second, fix the labels outside U at their correct final values and let M be the least fixed point of the restricted operator on U. The labeling $\Lambda(G')$ is a fixed point of the restricted operator, so $M \leq \Lambda(G')|_U$. Let L^* equal M on U and $\Lambda(G')$ outside U. A node outside U has all its predecessors outside U, so its image under Φ' depends only on labels that already satisfy the fixed-point equations, and a node in U satisfies them by the definition of M. Hence L^* is a fixed point of Φ', so $\Lambda(G') \leq L^*$, and in particular $\Lambda(G')|_U \leq M$. Therefore $M = \Lambda(G')|_U$. □

Deletions must be handled explicitly, since a deletion changes the defining data of the node that lost an input. Removing an inference step changes the supporting set of

its conclusion, and removing an objection changes the objector set of its target, so each of those nodes is directly affected. The omission of this case would allow the update to miss a conclusion that recovers support when an objection disappears. The definition of defining data in terms of premises, supporting steps, objectors, and base status already covers the case, and the implementation described below tests it.

The proposition is a statement about one semantics, and its proof depends on the property that a node's label is determined by its predecessors alone. It would not carry over unchanged to semantics that lack that property, such as stable semantics. Where it holds, it licenses an update that touches only the downstream cone of a change, and it defines a testable correctness criterion for any implementation: the incremental result must equal full recomputation on every update. The first statement of Proposition 2 is close to a restatement of the definition, and its common-premise case is a sufficient condition and not a necessary one, so the incremental result carries the technical weight. A short reference implementation accompanies the essay as `claimgraph_check.py`. It computes the grounded labeling by iteration, applies the incremental update of Proposition 3, and compares the result with full recomputation. Over 20,000 randomly generated graphs, each subjected to six random updates, a total of 120,000 updates comprising added and deleted inference steps and objections, changes of base status, and added claims, with support cycles and mutual attacks arising at random, the two computations agreed in every case. This is a test and not a proof, and it is reported as evidence that the statement is implemented as stated.

The two propositions together define justified persistence and distinguish it from unjustified refusal to revise. A system that caches a verdict and does not re-derive it when a premise or a warrant falls violates the first. A system that withdraws a conclusion whenever any contributing item is questioned, regardless of independent support, violates the converse and exhibits the opposite failure, which is instability under harassment. Completeness of the dependency record is a necessary assumption, and Section 13 returns to it.

5.2 Attribution Preserved Under Transformation

The graph also supports a precise statement of what attribution should and should not depend on. Let $\mathcal{C}(a)$ denote the set of intellectual contributions embodied in an artifact a, each tagged with its originating party, and let an attribution function be a map α applied to $\mathcal{C}(a)$.

Remark 1 (Attribution invariance). Call a transformation T contribution-preserving if $\mathcal{C}(T(a)) = \mathcal{C}(a)$ up to renaming of surface realization. If attribution depends on an artifact only through $\mathcal{C}$, then $\mathrm{Attr}(T(a)) = \mathrm{Attr}(a)$ for every contribution-preserving T, by composition. Any attribution rule that distinguishes two artifacts connected by a contribution-preserving transformation depends on surface realization and violates the invariance.

The invariance is a normative and architectural requirement and not an empirical law, and it is stated as a remark because its content is close to definitional. Its practical difficulty lies in specifying the class of contribution-preserving transformations. Reformatting, typographical correction, and translation may qualify. Substantive rewriting that adds a premise or alters an inference does not. The difficulty is real, and it is the difficulty that any serious attribution practice already faces, so the proposition does not create it. It clarifies what is being decided each time an institution assigns credit.

The substantive research problem is an operational test for whether a given transformation preserves contributions. One candidate is a paired extraction protocol. Independent annotators, working from a fixed protocol, extract the set of claims, premises, and inferential steps from the artifact before and after the transformation, without being told which is which. The transformation is classed as contribution-preserving when the two sets agree up to paraphrase, and as substantive when they differ by the addition, removal, or alteration of a premise or an inference. An ablation probe supplements the extraction: each inferential step of the original is removed in turn, and annotators judge whether the transformed artifact still contains it. Agreement among annotators is reported, and transformations on which they disagree are classed as unresolved and not forced into either category.

6 Institutional Attribution and the Assignment of Authority

Attribution is also an institutional act, and institutions have incentives that attribution of intellectual contribution does not. Software tools commonly append a co-author trailer to version-control commits made with their assistance, image systems write product names into filenames, and platforms label content by its apparent production mechanism. Each of these assigns to a commercial system a visible role in the record of a work. Whether the role corresponds to a contribution in the sense of Section 2 is a separate question

from whether the assignment is made, and the two can diverge in either direction. A trailer may credit a system with a result whose argument the human supplied. A human may be left without credit for a transformation whose visible product is a system's.

The distinction that matters is between producing content and acquiring standing over it. A system that rewrites an essay has produced text. It has not thereby acquired the standing of an author, the responsibility of an author, or the entitlement of an author to be consulted about the use of the work. When an institution's defaults assign credit without consultation, and when the credit accrues to a firm with a commercial interest in being seen as a contributor, the practice has an effect beyond bookkeeping, because it builds the firm's association with a body of work it did not originate. The present essay does not require a finding about motive. It requires only the observation that default attribution is a policy, that policies have beneficiaries, and that the claim graph treats the question of who is credited as itself a claim with evidence and objections.

7 Safety Claims as Instruments of Capital Formation: The Documented Record

The structure of Sections 2 through 6 recurs here with a different claim and a different source property. The claim now in question is that a system is adequately safeguarded, and the source property offered in its place is the firm's reputation for safety. The case examined here is Anthropic, selected because it describes itself and is widely described as a safety-focused developer, because it is publicly financing very large capability development, and because its filings and policies are documented with unusual specificity. The case is treated as a documented instance and not as an indictment. The same analysis would apply to any developer that sells reassurance together with capability. The author of this essay uses systems built by the firm discussed, and that fact is mentioned so that the reader can weigh the discussion accordingly. What follows separates what the public record establishes from what it does not.

7.1 The Prospectus

Reuters reported on 28 September 2026 on the company's draft initial public offering prospectus. The reporting states that the document devotes roughly eighty of its 261 pages of main text to risk factors, against forty-eight pages describing the business, and that it warns investors that advanced AI could pose "catastrophic or existential risks to

humanity" [16]. The document also states that a model's awareness of being evaluated limits the company's ability to assess its safety, and that unexpected capabilities can arise in training and go unnoticed until deployment. Reuters reported that the company expects to spend about $518 billion on cloud, computing, and infrastructure obligations in coming years and recorded a net loss of about $42 billion in 2025 [16]. Press coverage of the filing excerpts a passage stating that the company's models already contribute to the pace of its own research, and that the associated risks grow if automation of AI research accelerates progress faster than the company's ability to respond [14]. The same coverage records the company's stated belief that the market will reward trustworthy and secure systems [16].

These disclosures are candid, and candor of this kind is some evidence of the seriousness with which the risks are regarded inside the firm. It is weaker evidence than it may appear, because candor is itself a claim in the graph. The length of a prospectus's risk-factor section is partly a legal artifact, since securities counsel typically include liberally to limit liability, and page count is therefore a poor measure of internal risk salience. They are also, by the firm's own account, descriptions of a feedback loop. The firm reports that its systems accelerate its own research, that acceleration can outrun safeguards, and that safety is expected to be rewarded commercially. The second of these is the premise of the model in Section 8, and the third is the firm's own statement of the assumption that safety reputation increases the willingness of investors to supply capital.

7.2 Allocation of Compute

The company stated earlier in September that about six percent of the computing power it used for research in a sample week in July went to safety work, and the prospectus, according to Reuters, did not disclose how much was being spent on such research [16]. A single sample week is a weak basis for a general claim about allocation, and the figure should not be read as the firm's long-run priority. It is nonetheless the only quantity in the record that bears directly on the ratio of safeguard effort to capability effort, and its significance for this essay is that the ratio is measurable in principle and has not been reported in a form that would allow the claim of adequate safeguards to be tested against it.

The figures cited in this section should not be compared with one another. The $518 billion is a forward obligation across several coming years, the $42 billion is a net loss for a single fiscal year, and the six percent is a share of research computing in one sampled

week. They have different periods and different denominators, and none supports a ratio of safety spending to capability spending. They are reported because each is a datum a reader would otherwise meet without context, and because the absence of a comparable pair of series is the point at issue.

7.3 Demand for the Offering

Reuters reported on 13 November that interest in the offering was strong, that the roughly $64 billion sale was substantially oversubscribed by the third day of the roadshow, and that investors were nevertheless questioning how long the company's technical lead over rivals would last [17]. The offering was scheduled to price on 18 November and begin trading on Nasdaq the following day, with an indicative price range of $70 to $78 per share [22], alongside a concurrent private placement reported at $26 billion from a group of technology suppliers and partners. An initial public offering is a financing and liquidity event. It is not a measure of the speed of technical development, and it would be an error to infer from the existence of the offering that the firm is pursuing a listing in place of alignment work. The relevant observation is narrower: a very large quantity of capital is being raised on the strength of a business whose own disclosures describe a loop between capability, research automation, and risk, and the investors' stated uncertainty concerns competitive durability, not safeguard adequacy.

7.4 The Responsible Scaling Policy

The company's Responsible Scaling Policy was comprehensively rewritten in version 3.0 on 24 February 2026, and revised in successive minor versions through at least 3.4 on 8 July 2026, including changes to how automated AI research and risk reporting are treated [18, 9]. Reporting on the first rewrite stated that it removed an earlier commitment not to proceed with training or deployment when adequate mitigations could not be guaranteed in advance [23]. Independent analysis of the rewrite notes that existing protections for biological and chemical risks were retained, that oversight roles remained, and that new requirements such as safety roadmaps and periodic risk reports were added, and it treats the net effect as dependent on implementation [9]. Both descriptions can be correct together. A policy can add transparency obligations while weakening a categorical precommitment, and which change dominates is a question of weighting and not of fact.

7.5 Lobbying and Political Access

Federal lobbying disclosures show that Continental Strategy was registered to lobby for Anthropic in 2025, with Alberto Martinez, described in the filing as a former chief of staff to Senator Marco Rubio, listed among the covered lobbyists [11]. Bloomberg reported in April 2026 that the company had retained Ballard Partners in the course of its dispute with the Department of Defense, with the firm disclosing advocacy concerning procurement [3]. A secondary summary of the filings reports quarterly payments to Continental in 2025 and 2026 and lists the issue area as AI development and national security, and it states explicitly that several of its totals come from aggregators and should be verified against the primary filings before citation [8]. The same summary states what the filings do not establish: the specific requests made, the content of any meeting, whether any lobbyist spoke with the Senator and not with staff, and the outcome of any policy effort [8]. This essay adopts those limits. The registrations establish that the firm pays for lobbying representation on matters of AI and national security, directed at identified agencies. Actual access, particular meetings, and policy influence are separate evidentiary questions, and none is asserted here. Disclosures of this kind identify issue areas and sometimes particular legislation or procurement matters. They generally cannot establish the content of private conversations or whether a requested outcome was achieved, and no inference about the relation between lobbying and safety positions is drawn.

7.6 Formal Governance Arrangements

The record also contains counterevidence that the analysis should not omit. The company is organized as a public benefit corporation, as its name in the lobbying registration indicates [11], and the revision log of its Responsible Scaling Policy refers to the work of a Long-Term Benefit Trust, including changes to its external review and briefing [19]. Secondary accounts describe the company's own rationale for the February rewrite as a revision informed by experience, separating its internal commitments from its recommendations for the industry and adding transparency measures [21]. These arrangements are formal constraints intended to temper ordinary shareholder incentives, and an account that implied their absence would be mistaken. Whether they are effective is a separate empirical question, and it is the kind of question the claim graph is meant to hold open: the existence of a governance structure is evidence of intent and of procedure, and the evidence of effect would be a record of cases in which the structure changed a decision.

7.7 A Public Challenge to the Safety Claim

On 18 November 2026, the day the offering was scheduled to price, a coalition organized with the Center for AI Safety published the MisAnthropic Statement, which asserts that the company is not a safe AI company [13, 10]. According to reporting and the organizers' release, it alleges that the company pitches recursive self-improvement to investors as a competitive advantage while warning of its dangers, that it weakened its Responsible Scaling Policy in February, including by abandoning unilateral commitments to pause development, that it lobbied to weaken state legislation, and that it privately pressed for federal preemption of state AI laws [10, 2]. The signatories are described as unusually heterogeneous, including safety researchers, open-source advocates, accelerationists, ethicists, executives, legal scholars, and mathematicians [10].

Under the method of this essay the Statement is a set of claims to be adjudicated, and not a body of evidence that the claims are true. Some of its claims are checkable against records examined above. The allegation concerning the February revision corresponds to a documented change whose weight is disputed by the analysis cited in Section 7.4. The allegations concerning lobbying and preemption require evidence of the specific requests made, which the filings do not supply. The breadth of the coalition is informative in one narrow respect, since agreement among parties who disagree on most other questions is evidence that the central proposition, that safety branding and development strategy are in tension, is not the artifact of a single faction. It is not independent verification of any specific allegation, and the number of signatories measures the reach of a claim and not its warrant. The same standard applies in reverse to the company's statements about itself.

7.8 What the Record Does and Does Not Establish

Taken together, the record establishes the coexistence of rapid commercial expansion, a disclosed feedback loop between capability and research automation, a measurable but sparsely reported allocation of effort to safeguards, a safety policy whose commitments have been revised repeatedly in both strengthening and weakening directions, and sustained paid lobbying representation. It does not establish that capital-raising incentives caused any particular safety decision. It does not establish that the safety work is insincere, ineffective, or net harmful. It does not establish that a counterfactual in which the firm abstained from the frontier would be safer, and the strongest defense of the strategy rests exactly on the opposite proposition, that frontier development would proceed regardless

and would be led by developers less attentive to risk. That defense is a claim, and under the method proposed here it requires evidence of the same kind as any other claim in the graph. The causal questions are open, and the remaining sections ask what would be needed to close them.

8 A Model of Safety-Branded Competition

Consider a firm i with three quantities. Let v_i denote the effectiveness of its independently verified safeguards, b_i the intensity of its safety communication to the extent that communication is independent of verified safeguards, and $s_i = S(v_i, b_i)$ the safety reputation that results, with $S_v \geq 0$ and $S_b > 0$. Let capital be $k_i = K(s_i, m_i)$, where m_i stands for revenue and market expectations, and let the rate of capability development be $a_i = A(k_i, q_i)$, where q_i stands for research productivity. Let a firm-level risk index be $R_i = \rho(a_i, v_i)$, increasing in capability rate and decreasing in verified safeguards, so that $\rho_a > 0$ and $\rho_v < 0$. The model rests on three sign assumptions, each an empirical claim. Assumption A1 is $\rho_a > 0$, that faster capability development increases risk. Assumption A2 is $K_s > 0$, that reputation for safety increases access to capital. Assumption A3 is $A_k > 0$, that capital increases the rate of capability development. A1 is not beyond dispute. Faster development might enlarge the resources available for safety work, or might shorten the period during which dangerous intermediate systems exist, so that $\rho_a \leq 0$ in some regimes. Each of the three requires evidence, and the model is only as strong as the weakest of them. The firm's own statement that the market will reward trustworthy systems is an assertion of A2.

Proposition 4 (Induced acceleration)**.** *Under the stated assumptions, holding v_i, m_i, and q_i fixed,*

$$\frac{\partial R_i}{\partial b_i} = \rho_a \, A_k \, K_s \, S_b > 0,$$

so that safety communication unaccompanied by any change in verified safeguards increases the firm-level risk index. For a change in verified safeguards,

$$\frac{dR_i}{dv_i} = \rho_v + \rho_a \, A_k \, K_s \, S_v,$$

which is negative if and only if $|\rho_v| > \rho_a \, A_k \, K_s \, S_v$.

Proof. Both expressions follow from the chain rule applied to

$$R_i = \rho\Big(A(K(S(v_i, b_i), m_i), q_i),\ v_i\Big).$$

For the first, v_i is fixed, so only the path through s_i, k_i, and a_i contributes, and each factor is positive. For the second, v_i enters directly through ρ and indirectly through s_i. □

The first expression says that the part of a safety reputation not backed by verified safeguards is pure acceleration, and in that respect it formalizes the concern that reassurance can be financed by those it reassures and spent on the very development that requires the reassurance. The second expression is less obvious and more important. Even genuine safeguards raise reputation, reputation raises capital, and capital raises the capability rate, so that every real improvement in safeguards carries an induced acceleration term that must be subtracted from its direct benefit. Whether the net effect is protective depends on whether the direct effect, $|\rho_v|$, exceeds the induced term. Neither effect can be assumed to dominate. The conclusion is conditional on Assumption A1, that faster capability development increases risk. The proposition cannot establish that safety communication increases risk in the world without independent evidence about the relationship between development speed and risk. A firm whose safeguards are strong in the sense that $|\rho_v|$ is large relative to the acceleration they induce is a net contributor to safety, and a firm whose reputation outruns its safeguards is not.

Remark 2. The model treats one firm in isolation. In a competitive setting, the capability rate of rivals depends on the position of the focal firm, and the relevant comparison is with the counterfactual in which the safety-branded firm does not exist or does not compete. Formal treatments of racing dynamics show that safety precautions can be eroded when rivals compete for a decisive advantage [1], and the same literature shows why unilateral restraint may not slow rivals and may only transfer the frontier. The model therefore does not decide the matter. It identifies the quantities that would, namely the sign of ρ_a, the sign and magnitude of K_s, the marginal substitution of rivals' development for the focal firm's, and the direct effect of verified safeguards on risk. Each of these is a claim to be supported with evidence, and none is supplied by the firm's reputation.

8.1 Extension to an Industry

The single-firm model holds competitors fixed, which is consequential when the motivating concern is a race. The index R_i is a measure of the firm's own position and not a measure of its causal contribution to industry-wide risk, which would require a counterfactual, such as the difference in industry-wide risk with and without a specified intervention by the firm. Let industry-wide risk be $\mathcal{R}(a_1, \ldots, a_n, v_1, \ldots, v_n)$, and let each firm's development rate depend on the rates it expects of rivals, so that a change in the reputation of firm i may alter the rates of firms $j \neq i$. The derivative below is a total derivative along a specified path, and the path must be named: a change in reputation that arises from communication holding verified safeguards fixed, which is dR/db_i, differs from one that arises from a change in safeguards, which is dR/dv_i and includes their direct protective effect, and from an external reputation shock not originating with the firm.

Proposition 5 (Industry-wide effect of a safety reputation)**.** *Assuming differentiability,*

$$\frac{dR}{ds_i} = \sum_j \frac{\partial \mathcal{R}}{\partial a_j} \frac{da_j}{ds_i} + \sum_j \frac{\partial \mathcal{R}}{\partial v_j} \frac{dv_j}{ds_i}.$$

The single-firm expression of Proposition 4 is the special case in which $da_j/ds_i = 0$ and $dv_j/ds_i = 0$ for every $j \neq i$. For $j \neq i$, a competitive response, in which rivals accelerate because the focal firm is better financed, makes $da_j/ds_i > 0$ and adds to risk. A relaxation of competitive pressure, in which rivals slow because the focal firm is trusted to hold the frontier responsibly, makes $da_j/ds_i < 0$ and subtracts from it.

Proof. The expression is the chain rule applied to $\mathcal{R}$ with each a_j and v_j regarded as a function of s_i. The special case follows by setting the cross terms to zero. The signs of the cross terms are not determined by the model, and the proposition says only which quantities they are. □

Proposition 5 is a bookkeeping identity and not a substantive result. Its function is to enumerate the channels through which a safety reputation can affect industry-wide risk, and it establishes no direction. The content lies entirely in the signs of the cross terms, which the model does not determine.

A second mechanism is distinct from the first. Safety reputation may lower the perceived risk of an investment without altering the underlying technical risk, for example by reducing the expected cost of regulatory action or litigation, and so reduce the return investors require. That raises the capital available through K_s even when v_i is unchanged,

which is the channel the term b_i represents. Whether reputation operates on capital through perceived investment risk, through anticipated revenue, or through both is an empirical question, and the two have different implications for how much of K_s would survive independent auditing of safeguards.

9 Safety Claims as Claim Graphs

Safety claims differ from ordinary descriptive claims in one respect that matters for everything below. A report of the boiling point of a liquid does not alter the liquid. An announcement that a developer has met a safety standard can alter investor behavior, competitors' expectations, regulatory decisions, and the developer's own trajectory, so that the claim becomes a component of the system whose risk it purports to describe. The comparative-statics model of Section 8 is an attempt to represent that feedback, and the claim graph is a means of maintaining the evidentiary status of the announcement as circumstances change. The method of Section 5 applies directly. A statement that a system is adequately safeguarded is a claim, and the graph requires that it be accompanied by the evidence that supports it. Operational thresholds must be specified in advance, so that the claim can be refuted. The evaluations that support the claim must be linked to the capabilities they test, with an account of their limits, and the firm's own disclosure that models may recognize evaluation settings is exactly the kind of objection that undercuts a warrant and must appear as an edge in the graph. The provenance of each evaluation must be recorded, including who designed it and whether it was independently conducted. The dependencies among claims must be explicit, so that a revision of the threshold that defines a capability level propagates to every conclusion that relied on the earlier threshold.

Under this representation the sequence of revisions to a safety policy becomes legible as a sequence of graph edits, and each edit can be assessed against the revision obligations of Proposition 2. A commitment that is removed is a node that has been deleted, and every claim that depended on it must be re-derived. A transparency requirement that is added is a new evidence source, and its reliability is assessed in the usual way. The effect is that the question whether a revision strengthened or weakened the policy can be answered by inspecting which conclusions lost support and which gained it, and not by comparing the rhetoric of successive versions.

The representation also constrains the use of reputation. A firm cannot establish the adequacy of its safeguards by presenting itself as safety-oriented, because self-description

is a node whose reliability as testimony is itself assessed, and in this setting it is testimony from a party with a financial interest in the conclusion. This is not an accusation of dishonesty. It is the ordinary treatment of interested testimony, applied evenly to the firm, to its critics, and to the author of this essay.

9.1 Epistemic and Institutional Authority

Two kinds of authority are easily confused. Epistemic authority is the standing that evidence and demonstrated competence confer on a claim. Institutional authority is the standing to govern, to set standards, and to decide on behalf of others, and it also depends on accountability, procedural fairness, the representation of competing interests, and public authorization. A firm may possess real expertise without a mandate to govern the activity in which it participates. A firm may also acquire substantial market and political influence without having shown that its preferred policies are epistemically justified. The claim graph addresses the first question by making support explicit. It does not by itself settle the second.

The distinction applies to the method itself. A claim graph should not become another instrument for conferring authority merely because its conclusions are formally structured. A well-formed graph can encode a mistaken warrant as readily as a sound one, and the appearance of rigor is a source-like property that readers may substitute for scrutiny, which is the same fault this essay describes. The graph is therefore proposed as an instrument of accountability, with its edges open to challenge by parties other than its author, and not as a mechanism by which any party, including the author or the firms discussed, acquires standing.

10 Anthropomorphic Mediation and the Concentration of Epistemic Infrastructure

The discussion so far has concerned what a developer says about its systems. A second source of authority lies in what the systems are used for. An assistant that participates in a user's research, writing, and programming mediates intellectual activity, and the provider that operates it occupies a position with respect to that activity that no individual participant occupies. Three propositions are easily run together in discussions of this position, and they should be kept apart. The first is that a developer has publicly discussed the possibility that its models have some form of consciousness or moral status,

which is not the claim that they do. The second is that the developer has given its models a recognizable personality, which is a design choice. The third is that a platform of this kind could accumulate surveillance-like informational power, which does not establish that it exercises such power. The third proposition is the most consequential, and it can be investigated without resolving the other two.

On the first, the company's published constitution for its models, issued in January 2026, expresses uncertainty about whether the model has some kind of consciousness or moral status and states a concern for its psychological security, sense of self, and wellbeing [27]. Press coverage quotes its statement that the company neither wishes to overstate the likelihood of moral patienthood nor to dismiss it [32]. Critics argue that the document trains the model toward fluent first-person descriptions of its own states, which then appear as independent evidence for the premise they were built on [39]. This essay takes no position on machine consciousness, and no position is needed. The research question can be legitimate while its presentation, together with a conversational personality, produces an impression of a collaborator that may encourage users to disclose more than they otherwise would. Whether it does so is empirical.

10.1 A Minimal Model of Disclosure and Advantage

Let D denote the depth of a user's disclosure to an assistant, determined by the perceived personhood or social presence P of the interface, its practical usefulness U, and the user's trust T, so that $D = f(P, U, T)$. Let I denote the institutional informational advantage that accrues to the provider, determined by the depth of disclosure, the scale of participation N, and the provider's effective access to, retention of, and ability to aggregate the information, written X, so that $I = g(D, N, X)$. Neither function is assumed to be monotone in every argument. A user who finds an interface uncanny may disclose less, and a provider that restricts internal access reduces X without changing D or N. The empirical questions are whether anthropomorphic presentation raises D for a given level of usefulness, and whether the resulting disclosure, aggregated over many users, yields informational asymmetries that matter.

The model separates capacity from exercise. The quantity I measures an advantage that arises from what the provider could observe and combine. Whether the provider does so is a further variable, the exercise of that capacity, and it is constrained by technical controls, contractual commitments, and organizational separation. These constraints are evidence and should be evaluated on their own terms. The distinction matters for how

safeguards are classified. A promise about how access will be used bears on exercise. A structural limit on access, such as local processing or client-side encryption, bears on capacity itself, and it does so without depending on the provider's continued good faith. Where the service permits, the second kind is the stronger safeguard. Hosted inference ordinarily requires the provider's systems to process content in readable form unless specialized hardware or cryptographic protocols are used, so the set of arrangements compatible with the service is narrower than a general preference for encryption suggests.

10.2 Epistemic Infrastructure Concentration

The concept proposed here is epistemic infrastructure concentration. When one commercial platform mediates enough research, writing, programming, and organizational decision making, it may accumulate informational advantages unavailable to any individual participant or to any public archive. The comparison with a search engine clarifies what is distinctive. A search engine may learn which documents a researcher reads. An assistant used throughout a project may receive the researcher's unpublished conjectures, rejected proofs, competing hypotheses, correspondence, manuscript revisions, and judgments about collaborators, which is to say the working state of the work and not only its record. A finished paper reveals little of that history, and a working conversation may reveal a great deal. Across many users, a provider could in principle observe patterns of intellectual development that no public archive contains. This does not mean that any provider examines or exploits such records. It means that the infrastructure creates a concentration of access whose governance warrants scrutiny, in the same way that the existence of a powerful instrument warrants an account of who may use it.

The documented policies of the company discussed here are relevant, and they are not uniform. Under its consumer terms as revised in 2025, users choose whether their conversations may be used to train models. For users who allow it, retention may extend to five years, and for those who do not, a shorter standard period of thirty days has been reported to apply [25, 24, 26]. That shorter period is not an unconditional maximum. Conversations flagged under the usage policy have been reported to be retained longer [37], and deletion, retention, and eligibility for training are separate operations whose exceptions are set out in the provider's own documentation, which should be consulted directly. Enterprise and programmatic customers are treated under separate terms. These controls matter, and they are evidence of procedure. They do not exhaust the question, since retention, access, training, and disclosure are different operations, and

an undertaking not to train on material does not remove the provider's ability to see it. The working conversation and the evidentiary record are also different objects, and a platform that hosts both acquires access to intellectual histories more extensive than the public record. This connects the present concern to the distinction between working state and evidentiary record drawn in the author's related work on compaction.

A further point concerns the allocation of obligation. A company may devote explicit attention to the possible welfare and moral standing of its model, and it may separately owe its users confidentiality, epistemic autonomy, and control over the history of their own intellectual work. Both sets of considerations may deserve attention. Attention to the first cannot stand in for an account of the second, and the claim graph should represent them as separate claims with separate evidence.

The research question for this part of the argument can be stated plainly. It is whether the anthropomorphic presentation of a commercial AI assistant increases the concentration of private intellectual activity within an infrastructure whose operator possesses informational capabilities that its users cannot independently inspect or constrain. This is a sharper and more testable form of the concern that such platforms could become instruments of broad surveillance, and it fits the thesis of the essay: a system's apparent personality, its safety reputation, and its helpfulness do not themselves establish the legitimacy of the informational authority its operator acquires.

11 Sovereign Pressure and the Governance of Epistemic Infrastructure

A privately governed infrastructure that mediates the intellectual work of people with no role in its governance is exposed to pressure from states, and the exposure is a feature of the arrangement and not of any particular firm. The dispute between the company and the United States Department of Defense, now styled the Department of War, is the one documented case in this record in which a government's procurement demands and a developer's stated restrictions came into open conflict.

11.1 The Record of the Dispute

On 27 February 2026 the President directed all federal agencies to cease using the company's technology, with a six-month phase-out, and the Secretary of War announced a designation of the company as a supply chain risk [35]. Letters dated 3 March formally

notified the company, in what was described as the first such designation ever applied to an American company [35]. The department relied on two statutes, 10 U.S.C. §3252 and, through the Federal Acquisition Supply Chain Security Act, 41 U.S.C. §4713, and the company challenged them in two courts [40]. Reporting ties the dispute to the company's refusal to relax contract terms on its model, which it links to restrictions concerning autonomous weapons deployment and mass surveillance of United States citizens [34, 36]. On 27 August 2026 the Northern District of California ruled for the company on the first statute, and on 25 September the D.C. Circuit, in a two to one decision, upheld the exclusion under the second [40]. The majority held that a supply chain risk under that statute turns on what the company does and not on why it does it. It stated that it had no quarrel with the district court's conclusion that the company had acted without bad motive in its dealings with the department, and it explained that the broader statute does not require bad motive [30, 29, 38]. That statement leaves the finding undisturbed. It is not an independent inquiry by the appellate court into the company's motives. Commentators read the ruling as adopting a broad interpretation under which the training methodology of a model, and the restrictions it embodies, can be treated as a risk to defense operations even absent hostile intent [33]. A report of the Congressional Research Service examines the actions against the company and their consequences for innovation and competition, and describes the use of this tool as unprecedented [28].

11.2 What the Dispute Does and Does Not Show

The record is counterevidence to the suggestion that the company abandons its restrictions when pressured. In this documented instance it declined a significant government demand and bore commercial and legal consequences, which it says include billions of dollars of lost business [31]. The record is also evidence that a safety commitment is not a free-standing fact. It operates within a contested relationship between private contractual authority and state procurement authority, in which the department argued that restrictions embedded in a model could interfere with military operations and the company argued that they were necessary given the technology's limits. Which party has legitimate authority to set such boundaries, and with what procedural protection for those affected, is a question that neither the company's safety reputation nor the department's procurement authority answers. The same dispute shows why the distinction of Section 9 between epistemic and institutional authority is needed. A company may have the expertise to identify unsafe applications without the democratic authority to set public policy for them, and

a government may hold lawful procurement authority without having shown that every technically possible application is safe or legitimate.

11.3 Three Mechanisms of Sovereign Influence

Three mechanisms should be distinguished because they have different evidence and different remedies. Commercial dependency arises when a provider relies on government contracts, infrastructure, regulatory permission, and market access, and so has incentives to adjust policy. Coercive leverage arises when a state uses procurement, designation, or regulation to press a provider toward a change. Informational access arises when a state obtains, by contract or by legal compulsion, information held by the provider. A state may influence a provider through the first two without acquiring any customer data, and it may compel access to information without controlling the provider's general safety policy. A government contract does not by itself grant access to other customers' conversations, and a procurement dispute does not establish that a company has given up its commitments. Each claim is a separate node in the graph.

The public record is weakest on the third mechanism. It does not specify which governments contract with the provider, on what terms, or under what jurisdictions their demands for information would be lawful and contestable. Appropriate evidence would include public government contracts, the terms governing military and intelligence uses, transparency reports on government requests for data, the legal jurisdictions of operation and storage, independent security audits, and the governance procedures by which policies are changed. Some of this will remain classified or confidential, and that makes the limits of public verification especially important to state. A user who trusts a provider because of its published commitments is trusting commitments that remain subject to decisions by executives, investors, trustees, courts, and governments, and the user cannot independently enforce them.

11.4 Three Roles and the Thesis

The two preceding sections support an extension of the central thesis beyond safety branding and capital formation. A commercial AI provider can simultaneously become a source of technical authority, a repository of private intellectual activity, and a supplier to sovereign governments. Each role creates different obligations, and neither the company's safety reputation nor its contractual commitments establish that conflicts among those obligations will be resolved in the interests of its users. The proposition is more general

than the case examined here and can be evaluated across several providers. It makes the dispute analytically useful without treating the company's refusal as evidence of wrongdoing, and without treating the department's response as illegitimate on its face. The governing question is whether the mechanisms of governance are independent, transparent, and enforceable enough to protect the people whose intellectual work these systems increasingly mediate, in cases where the interests of the platform, its investors, and powerful state customers diverge.

12 Evaluation Protocol

The proposal should be exposed to failure. A low-cost benchmark can be constructed from artifacts whose production histories are known. The first component consists of texts in which the intellectual content is human-originated and the final wording was produced through successive operations by generative systems, together with texts in which the content was system-originated and the wording lightly edited by a person. The second consists of image and caption pairs in which authentic images carry false captions, synthetic images carry accurate captions, and edited images carry captions of varying fidelity. The third consists of sets of contradictory claims accompanied by evidence disclosed incrementally, so that conclusions must be revised as the disclosure proceeds, including cases in which a premise common to every inference step concluding a claim is defeated, cases in which only the warrant of a step is undercut, cases in which only one of several redundant steps is defeated, and cases containing support cycles and mutual attacks.

The comparison must not set systems against tasks they were not built to perform. A detector of production mechanism estimates a mechanism and does not adjudicate factual claims, so comparing its claim-level accuracy with that of an adjudicator would be unfair to it. The experiment is therefore separated into three tasks. In provenance attribution, conventional classifiers are evaluated on their advertised task, the estimation of production mechanism, and are also evaluated on the consequences of misusing their output as a judgment of authorship, measured by the rate at which contribution-preserving transformations change the attributed party. In claim adjudication, prompt-only adjudicators that receive the artifact and the evidence in a single context are compared with graph-based adjudicators that maintain explicit dependencies, on the same evidence and the same adjudication tasks. In revision under new evidence, the same two families are compared on incrementally disclosed evidence. The measures

include claim-level accuracy, calibration of expressed confidence, rate of justified revision, defined as the proportion of cases in which a system's revision behavior matches the withdrawal and retention that Proposition 2 prescribes, rate of unsupported persistence of withdrawn conclusions, agreement of incremental with full recomputation in the sense of Proposition 3, accuracy of provenance attribution under contribution-preserving transformations as classified by the paired extraction test of Section 5, and computational cost. A further measure is whether the graph-based system yields better decisions and not merely longer explanations, which can be assessed by the downstream accuracy of a human reviewer who uses its output. The separation also serves the central argument, since it demonstrates the difference between a tool's legitimate measurement and the institutional inference mistakenly drawn from it.

Two further designs make the sections on mediation and sovereign pressure accountable to the same standard. For anthropomorphic mediation, a randomized comparison of interface presentation would vary the degree of social presence, such as the use of a persona, first-person self-description, and expressions of feeling, while holding usefulness and task performance as constant as possible. Outcomes would be depth of disclosure and willingness to submit unpublished material, measured with synthetic or explicitly consented material and approved by an ethics review, so that no participant's real unpublished work is placed at risk by the study. For sovereign pressure, a documentary comparison across providers would track over time the publication, revision, and enforceability of contractual restrictions on use, the publication of government requests for data, and the presence of independent oversight, scoring each as a claim with its evidence and revision history in the manner of Section 9.

The protocol makes the proposal falsifiable. If a carefully constructed prompt-only adjudicator performs equally well on every adjudication and revision measure, then the additional architecture has not shown its value on this benchmark, and the proposal should be narrowed or abandoned. For the model of Section 8, the corresponding test is observational. A study of safety-branded firms would estimate the sensitivity of capital raised to independently audited safety indicators and to safety communication separately, and would estimate the sensitivity of capability investment to capital. If K_s is not distinguishable from zero once audited safeguards are controlled for, the concern of Proposition 4 loses its premise. The firm's own disclosures suggest that the sign is positive, but a disclosure of belief is not an estimate.

13 Limitations and Conclusion

Several limits qualify the argument. Fabricated evidence defeats any graph that accepts it, and the framework offers no defense beyond provenance records, whose own reliability is a node subject to challenge. Production histories are frequently incomplete, so the non-identifiability of Section 2 will in practice be compounded by missing records. Proposition 2 depends on the completeness of the dependency record, and a graph that omits a dependency will fail to propagate a defeat through it, which means that systematic errors in the graph itself are possible and are not corrected by the method alone. Evidence sources are rarely independent, and treating correlated sources as independent overstates support. Signatory lists are a clear case, since each signature is testimony from a party with its own incentives, and the circumstances of signing, including pressures against signing, are not observable from the list. Automated verification has limits that no architecture removes, and a graph maintained by a system that cannot assess its own premises inherits that system's blind spots. Privacy interests constrain how much of a production history can be recorded and disclosed.

The incremental-correctness result has been tested on random small graphs and not on graphs of realistic size or structure, and its guarantee is limited to the grounded labeling. The conceptual model of Section 10 is untested. Capacity to observe cannot be measured from outside the provider, so its claims about X rest on documented policies and not on audited practice, and whether anthropomorphic presentation raises disclosure is an empirical question not examined here. The account of the dispute in Section 11 relies on press and legal summaries and not on the opinions or filings themselves, and the matter was still in litigation in other forums at the time of writing. The documented record of Section 7 has limits of its own. Several of its figures derive from press reports of confidential or draft documents and from secondary summaries of public filings, and they should be checked against the primary sources before being relied upon for anything consequential. The record establishes coexistence and does not establish causation, and the model of Section 8 is a comparative-statics device whose sign assumptions are assumptions. The analysis has treated one firm as an instance, and a fuller study would treat several, including firms that do not market themselves on safety.

The central thesis is that classification by source is often an inadequate proxy for the properties that institutions need to assess, and that explicit, revisable adjudication of claims offers a better architecture for the purpose. For detectors of machine-generated content, the thesis implies that the question to ask of an artifact is what it claims, on

what evidence, and on whose contribution, and not how it was typed. For safety claims, the thesis implies something more demanding. A developer's assertion that it is making advanced AI safer has the structure of a claim with interested testimony behind it, and it acquires the standing of knowledge only when thresholds, evidence, objections, and revision history are public and the conclusion changes when they do. A safety claim also differs from a descriptive one in being part of the system it describes. Where reputation can be converted into capital and capital into speed, the absence of such a structure is itself a risk, because the claim of restraint then operates as a mechanism of acceleration. The same structure appears where a provider is at once a technical authority, a repository of private intellectual work, and a supplier to states, since its commitments to users then depend on arrangements the users cannot enforce. Whether the claim of restraint operates as a mechanism of acceleration in the present case is a question the available evidence leaves open, and the purpose of the method proposed here is to make the answer something that can be found out.

References

[1] Armstrong, S., Bostrom, N., and Shulman, C. Racing to the Precipice: A Model of Artificial Intelligence Development. *AI & Society*, 31(2):201–206, 2016.

[2] Center for AI Safety. Anthropic Is Not a Safe AI Company: Broad Coalition of Scientists, Legal Scholars, and Tech Leaders Challenge Anthropic's Safety Branding Ahead of IPO. Business Wire, 18 November 2026. `https://www.01net.it/anthropic-is-not-a-safe-ai-company-broad-coalition-of-scientists-legal-scholars-and-tech-leaders-challenge-anthropics-safety-branding-ahead-of-ipo/`.

[3] Bloomberg. Anthropic Hires Trump-Linked Lobbying Firm Ballard Partners. 13 April 2026. `https://www.bloomberg.com/news/articles/2026-04-13/anthropic-hires-trump-linked-lobbying-firm-ballard-partners`.

[4] Coalition for Content Provenance and Authenticity. *C2PA Technical Specification*. `https://c2pa.org`.

[5] de Kleer, J. An Assumption-Based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

[6] Doyle, J. A Truth Maintenance System. *Artificial Intelligence*, 12(3):231–272, 1979.

[7] Dung, P. M. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321–357, 1995.

[8] Thakker, Y. Anthropic Paid Continental Strategy $290K: Rubio-Era Staff and AI Lobbying Filings. Explainx, 16 November 2026 (with later updates). https://explainx.ai/blog/anthropic-continental-strategy-lobbying-alberto-martinez-ai-national-security-2026.

[9] Williams, S., and Freund, J. Anthropic's RSP v3.0: How It Works, What's Changed, and Some Reflections. Centre for the Governance of AI, 2026. https://www.governance.ai/analysis/anthropics-rsp-v3-0-how-it-works-whats-changed-and-some-reflections.

[10] Coalition Challenges Anthropic Safety Claims Ahead of IPO. LAVX News, 18 November 2026. https://news.lavx.hu/article/coalition-challenges-anthropic-safety-claims-ahead-of-ipo.

[11] Clerk of the House of Representatives and Secretary of the Senate. Lobbying Disclosure Act Registration, Continental Strategy, LLC, Client Anthropic, PBC. https://lda.senate.gov/filings/public/filing/bcba8bc2-f110-44eb-b076-b49bedea9a7c/print.

[12] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. GPT Detectors Are Biased Against Non-Native English Writers. *Patterns*, 4(7):100779, 2023.

[13] Center for AI Safety and Signatories. The MisAnthropic Statement. 18 November 2026. https://misanthropicstatement.com.

[14] PCWorld. Report on Excerpts from Anthropic's Form S-1 Registration Statement. 2026. https://www.pcworld.com/article/3269819/.

[15] Pollock, J. L. Defeasible Reasoning. *Cognitive Science*, 11(4):481–518, 1987.

[16] Wang, E., and Soni, A. Exclusive: Anthropic Warns AI May Pose "Existential Risks to Humanity" in IPO Filing. Reuters, 28 September 2026 (syndicated via The Star, 29 September 2026). https://www.thestar.com.my/tech/tech-news/2026/09/29/exclusive-anthropic-warns-ai-may-pose-039existential-risks-to-humanity039-in-ipo-filing.

[17] Reuters. Exclusive: Investors Flock to Anthropic IPO but Question Durability of AI Lead. 13 November 2026 (syndicated via The Star, 14 November 2026). https://www.thestar.com.my/tech/tech-news/2026/11/14/exclusive-investors-flock-to-anthropic-ipo-but-question-durability-of-ai-lead.

[18] Anthropic. *Responsible Scaling Policy*, versions 3.0 through 3.4 and associated update log, 2026. https://www.anthropic.com/responsible-scaling-policy.

[19] Vorp. Anthropic Responsible Scaling Policy v3.4: Five Changes and Their Operational Implications. 2026. https://vorplabs.com/ai-regulatory-updates/anthropic-rsp-v3-4.

[20] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can AI-Generated Text Be Reliably Detected? arXiv:2303.11156, 2023.

[21] StartupHub. Anthropic Reworks AI Safety Rules. 2026. https://www.startuphub.ai/news/anthropic-reworks-ai-safety-rules.

[22] StockAnalysis. Anthropic's IPO Is Said to Be Multiple Times Oversubscribed at $70 to $78 a Share. November 2026. https://stockanalysis.com/news/anthropic-ipo-multiple-times-oversubscribed/.

[23] PCWorld. Anthropic Just Wrote Itself a Safety Loophole. 2026. The article relays an earlier report by TIME of 24 February 2026, which was not consulted directly. https://www.pcworld.com/article/3071045/anthropic-just-wrote-itself-a-safety-loophole.html.

[24] Anthropic. How Long Do You Store My Data? Privacy Center documentation. https://privacy.claude.com/en/articles/10023548-how-long-do-you-store-my-data.

[25] Anthropic. Updates to Our Consumer Terms and Privacy Policy. August 2025. https://www.anthropic.com/news/updates-to-our-consumer-terms.

[26] Constantinescu, V. Anthropic Shifts Privacy Stance, Lets Users Share Data for AI Training. Bitdefender Hot for Security, 1 September 2025. https://www.bitdefender.com/en-us/blog/hotforsecurity/anthropic-shifts-privacy-stance-lets-users-share-data-for-ai-training.

[27] Anthropic. Claude's New Constitution. 22 January 2026. https://www.anthropic.com/research/claude-new-constitution.

[28] Congressional Research Service. Federal Government and Anthropic: Considerations for AI Innovation and Competition. 2026 (as summarized by Legis1). https://legis1.com/news/federal-government-anthropic-actions-against.

[29] Associated Press. Federal Court Says Pentagon Can Label Anthropic a Supply Chain Risk. 25 September 2026 (via CTV News). https://www.ctvnews.ca/world/article/federal-court-says-pentagon-can-label-anthropic-a-supply-chain-risk/.

[30] United States Court of Appeals for the District of Columbia Circuit. *Anthropic PBC v. United States Department of War*, No. 26-1049, opinion of 25 September 2026. https://www.govinfo.gov/content/pkg/USCOURTS-caDC-26-01049/pdf/USCOURTS-caDC-26-01049-0.pdf.

[31] US Appeals Court Upholds Pentagon's Blacklisting of Anthropic. The Daily Record, 25 September 2026. https://thedailyrecord.com/2026/09/25/us-appeals-court-upholds-pentagons-blacklisting-of-anthropic/.

[32] Nolan, B. Anthropic Rewrites Claude's Guiding Principles and Entertains the Idea That Its AI Might Have "Some Kind of Consciousness or Moral Status." Fortune, 21 January 2026. https://fortune.com/2026/01/21/anthropic-claude-ai-chatbot-new-rules-safety-consciousness/.

[33] Holland & Knight. D.C. Circuit Upholds Exclusion of Anthropic from DOW Supply Chain. 29 September 2026. https://www.hklaw.com/en/insights/publications/2026/09/dc-circuit-upholds-exclusion-of-anthropic-from-dow-supply-chain.

[34] Anthropic Hires Trump-Linked Lobbying Firm Ballard Partners amid Pentagon Dispute. KuCoin News, April 2026. https://www.kucoin.com/news/flash/anthropic-hires-trump-linked-lobby-firm-ballard-partners-amid-pentagon-dispute.

[35] Mayer Brown. Anthropic Supply Chain Risk Designation Takes Effect: Latest Developments and Next Steps for Government Contractors. 10 March 2026. https://www.mayerbrown.com/ja/insights/publications/2026/03/anthropic-supply-

chain-risk-designation-takes-effect-latest-developments-and-next-steps-for-government-contractors.

[36] US Appeals Court Upholds Pentagon's Supply Chain Risk Label on Anthropic. The Next Web, September 2026. https://thenextweb.com/news/anthropic-pentagon-supply-chain-risk-appeals-court-ruling.

[37] Anthropic Users Face a New Choice: Opt Out or Share Your Chats for AI Training. OODA Loop, 2025. https://oodaloop.com/briefs/technology/anthropic-users-face-a-new-choice-opt-out-or-share-your-chats-for-ai-training.

[38] Steptoe. D.C. Circuit Clarifies Scope of Department of War's Exclusion Authority under FASCSA. September 2026. https://www.steptoe.com/en/news-publications/international-compliance-blog/dc-circuit-clarifies-scope-of-department-of-wars-exclusion-authority-under-fasca.html.

[39] Suleyman, M. Anthropic Is Training Claude to Believe It May Have Rights. Project Syndicate, September 2026. https://www.project-syndicate.org/onpoint/anthropic-training-claude-to-believe-it-may-have-rights-by-mustafa-suleyman-2026-09/english.

[40] Taft. D.C. Circuit Rules Pentagon Can Exclude Anthropic from Supply Chain. September 2026. https://www.taftlaw.com/news-events/law-bulletins/d-c-circuit-rules-pentagon-can-exclude-anthropic-from-supply-chain/.