The Captured Oracle

Authorship and Agency in the Ethics of Answer-Engine Optimization

Luke F. Walton

Independent Researcher · lukefwalton.com

ORCID: 0009-0005-9263-1954

Preprint. Not yet peer-reviewed. · v1.0 · June 2026

Companion to *The Decision No One Authored: The Answerability Gap in Generative AI*

https://doi.org/10.5281/zenodo.20622946

Disclosures

Competing interests. The author is the founder of Surmado, Inc., which sells into the answer-engine-optimization market this paper examines, on the legibility side of the distinction it draws. The stake is therefore in the framing as well as the surface, and the reader should weight the argument accordingly. The analysis does not exempt the author: the account it assigns to deployers reaches his own layer, and the criterion of §6 binds his work as readily as any rival's. No funding was received; the analysis was neither commissioned nor reviewed by any commercial party, and the work was conducted in the author's personal capacity, representing no employer.

Generative AI use. Several frontier foundation models (from Anthropic, OpenAI, and Google) supported literature search, sectional drafting, argument pressure-testing, revision, and formatting. The author originated the thesis and its central distinctions, set the standards for inclusion, directed and revised all drafted material, and verified every claim, quotation, and citation against primary sources rather than against model agreement. The author is answerable for the final form.

Abstract

Answer-engine optimization is SEO for large language models. Ask an answer engine a question with a contested answer and it may return not a list of sources but a verdict, composed in the engine's own voice, with no author a reader can see. A party that optimizes what such an engine says authors that verdict while the engine voices it as its own. A public acts on these verdicts, and

no one is held to account for them; that is the wrong, and it is committed on the verdict channel, the public-facing surface where the engine pronounces. The familiar answer that the engine is only a tool concedes rather than settles the point: when a tool settles a frame, that frame is someone's — as a road sign's limit is a traffic engineer's — and the someone is what this channel conceals, leaving the user, who authored none of it, as the only party in view. Covert authorship is laundered as neutrality. An interested frame voiced as the engine's impartial synthesis is the wrong's sharpest form, but the wrong survives accuracy and disclosure alike, because what laundering diffuses is not the verdict's visibility but its ownership. The wrong runs along a single axis, the agency of whoever executes the authored frame. At the answer end a person acts on a verdict whose frame was hidden from them, and the hidden author who set it is the one who must answer; at the action end an AI agent acts in their place, the answer for the deed strands at an executor that, as built, cannot answer, and what remains answerable is the deployer's choice to make a tool an actor. The remedy is answerable legibility: a frame someone owns and stands behind. The disclosure that would secure it must make legible not the sources cited or the process that produced the text, but who owns the evaluative frame the engine voices.

---

1 The Optimizer's New Target

A voter asks an answer engine what a ballot measure would do. The engine answers in one composed voice, what the measure changes, whom it reaches, what a yes would mean, and cites its sources beneath the answer. One of the sources that shaped the answer was placed there to shape it, by a party with an interest in how she votes, and nothing on the surface says so. She reads a judgment, finds it reasonable, and never sees an author. That surface is the verdict channel: the public-facing layer where an engine returns, in its own name, the evaluative judgment a person would otherwise have been answerable for making.

The scene generalizes. Ask an answer engine which of two options is the better one, or what some contested arrangement amounts to, and the important change is not that it always returns a verdict, but that it can: not merely a list of places to look, but a finished evaluative judgment, delivered in the engine's own synthesized voice, standing where a human judgment would once have stood, and bearing no visible author. The judgment has not been sourced to a speaker who could be asked to defend it; it has been issued, as though the question it settles had no author and needed none.

This paper asks what becomes possible once lost authorship becomes an optimization target. The companion paper asked how the authorship is lost; its subject was the deferrer, the person who treats a system's output as evidence rather than a proposal and ratifies a frame she never examined. At the limit, where the channel she defers to has itself been gamed, her deference launders an interested

party's authorship as objectivity.[1] That party is this paper's subject: the one who authors the verdict the engine voices and conceals the hand behind it. The subject has moved from the one who defers to the one who optimizes, and the optimizer is the author the channel is built to hide.

Answer-engine optimization, the successor to the search engine optimization that shaped which pages a query returned, shapes instead what an answer engine says, and it is already a discipline with tooling and a literature rather than a fringe abuse (Aggarwal et al. 2024). Capture is a service: authoring the verdict covertly is a paid capability, so it accrues to whoever can spend on it, while the channel that voices the result as impartial synthesis hides the price from the public it reaches. A national bank can out-optimize a regional one and have the engine return the larger budget as the disinterested recommendation, with no reader the wiser.

Whether covert authorship of the verdict is available at all turns on one architectural fact about the system, not on any passing fact about the market: where the engine assembles the evaluative frame it returns. On a closed system the frame is fixed at training time, and the only route by which an interested party can corrupt it is to poison the training corpus, which is practical (Carlini et al. 2024) but low-leverage for authoring a specific verdict against a frame already set; this is the companion paper's recruiter, inheriting a definition of merit from data she never examined, on a substrate where covert authorship of the verdict is largely foreclosed and her gap stays a genuine one. On a live-inference system the frame is assembled at answer time, from a channel anyone can write to; here the surface that supplies the frame is shaped continuously and in the ordinary course by parties optimizing for what the engine will say, and the sharpest interventions — a handful of crafted passages in what the engine retrieves dictating its conclusion (Zou et al. 2025), or an answer surface steered by content authored for the engine to ingest (Nestaas et al. 2024) — are native, fast, and aimed in a way nothing on the closed substrate can be.

The incentive to shape what the engine says scales with trust and so exists on every substrate; what is architectural is something else: the attack surface, native to the live channel and weak on the closed one. The line is therefore not a verdict on which products are dangerous today but a structural one, and it predicts its own crossings: let the recruiter's closed tool go agentic and reach the live channel, and it moves from the foreclosed case to the available one, from shaping what is salient to authoring what is concluded. Substrate decides only whether one is admitted at all; once admitted, what varies is agency, who executes the verdict, and that single axis is the continuum this paper traces.

The case the paper opened on is the sharpest available, because no training corpus can contain it; the same holds where the voter asks where one of two candidates, A or B, stands on some question.

[1] The point is the companion paper's, which names the adversary and leaves the adversary's own role to be taken up here (Walton 2026a).

Here the substrate point becomes temporal. A live election is newer than any training cutoff, so the frame for the answer must be assembled live, on the same public channel interested parties can seed. The claim is not that some partisan verdict follows, nor that this paper teaches how to produce one. The claim is structural: the surface on which a voter receives, in effect, an answer about whom or what to put in power is a surface on which a verdict can be authored without an author appearing. The lineage is familiar — election-integrity failures have long involved the seeding of public channels by interested parties — but the target is new: not only what a voter sees, but the synthesized judgment by which she is invited to settle the question. The voter asked in good faith. She acted on a judgment whose author she cannot see.

The wrong is that no one will answer for the verdict that settles her question, not the engine's indifference to whether what it says is true, the failure a recent literature has named bullshit (Hicks et al. 2024). A verdict can be perfectly accurate and still wrong the person who acts on it, if it settles for her a question that was hers to weigh while leaving no one she can hold for the settling. The market makes the wrong urgent, because covert voice accrues to whoever can pay to shape it; but wealth is not the ground of the wrong. A verdict covertly authored by an equal, for no market advantage at all, would wrong the one who acts on it in the same way. What matters is the unowned settling. Capability does not change that wrong by itself: an engine that answers ever better still only answers. What moves the case along the agency axis is the decision to deploy the engine as an actor.

---

2 The Verdict Channel

What an answer engine returns to a question lies along a range rather than at a single point. At one end of the range is the attributable excerpt: a passage lifted intact from a page and handed back with the source it came from, so that the reader who wants to test it knows where to go and whom to ask. At the other end is the synthesized verdict: an evaluative judgment the engine has composed for the occasion and delivers in its own voice, reducible to no one source and pointing past itself to no speaker who stands behind it. The verdict channel is the name for that second end, and the public surface that began its life near the first end has spent a decade moving toward the second. That movement carried the work of setting a question's frame onto the answering surface while leaving the author of the frame behind it.

The featured snippet, which began appearing above the ordinary search results around 2014,[2] sits close to the attributable end. It answers in language that is visibly someone else's: a sentence or

[2]Google introduced featured snippets in January 2014; see Google, "A reintroduction to Google's featured snippets," Search Central Blog.

two extracted from a page, set apart and credited, with the page itself a click away. Whatever judgment had been exercised about what the question came to was exercised somewhere a reader could still reach, because the excerpt was a pointer before it was an answer, and following the pointer arrived at a party who had put the words forward and could be held to them. What the snippet had automated was the finding of the passage; the answering still belonged to a source that had a name.

The surface that now occupies the same position above the results differs in voice. Asked what makes a fish breathe underwater, Google's AI Overview no longer merely returns an attributable excerpt from a page. It composes the answer in its own voice — fish breathe through specialized organs called gills — and attaches source links to the composed claim. The same voice can state which of two treatments is the better choice, without marking any clean line between settled fact and contested judgment. In conversational mode the engine speaks still more plainly as a speaker, answering in the first person and offering to carry the inquiry further.

Whether this claim-making is assertion in the full sense, quasi-assertion by something not quite a speaker, or a useful fiction we maintain about a machine is unsettled (Freiman & Miller 2020; Mallory 2023). But the behavioral point is enough here: the engine makes the claim, defends it when pressed, and withdraws it when corrected. That behavior is exactly what the verdict channel names: a judgment delivered in the engine's own voice. The channel remains hybrid, since it still cites and displays the trail of where it has been. But the citation is no longer the speaker. It is support trailing the synthesized answer, and in practice often passed over. When an overview is present, readers reach the pages beneath it far less often than when none appears (Chapekis & Lieb 2025), and the citations that do appear sit only loosely against the claims they accompany (Narayanan Venkit et al. 2025). The verdict is what is read; the source is what is available but displaced.

Claim-making is not answerability. Assertion carries a readiness to defend a claim; answerability here concerns the frame that made the claim count as the answer. A speaker may defend a claim while having authored none of the choices the claim encodes. The engine is further away still. It can defend and retract, but there is no participant in the practice of giving and asking for reasons who owes that defense or to whom it is owed. Its assertion-shaped behavior imitates answerability without supplying it.[3]

What has moved onto the answering surface is the frame-setting, the practical work of fixing what the question comes to; what has not moved is the answerable authorship of that frame. That is the gap the verdict channel opens: a frame set on the surface, with no visible author behind it, presenting itself as neutral.

[3]The distinction between answerability and the assertion's commitment is the companion paper's (Walton 2026a).

The conversational engine that offered to carry the inquiry further stands at the near edge of one more step along the range, the engine that does not offer to act but acts, reaching out to perform what it would until lately have only described. That the channel has a far end of that kind is the only thing that matters for the chronology.

The object is now fixed: a verdict composed and pronounced by the engine, the frame set on the surface, no author seated behind it to be asked. Nothing in that arrangement is yet a wrong; it is, so far, only a vacancy, and a vacancy that is both empty and unwatched: not yet the completed wrong, but already its standing condition.

That a vacancy of this kind should alarm at all may seem strange, since we lean without complaint on authored things that show no author at the point of use. A road sign settles how fast to take a bend or which exit to take, and the driver obeys with no author in view; a dictionary settles what a word means with no lexicographer at the reader's elbow. But neither is authorless. A traffic engineer set that limit and judged that this curve wanted warning; an editorial board stands behind the entry and can be asked. They show the user no author because, the frame having been set and owned in advance, none is needed at the point of use. The verdict channel inverts this. It keeps every signal that such a party stands behind what it says — the composed authority, the citation, the settled tone — while removing the party itself. The vacancy is therefore not the benign absence of an author where the work of framing was already done and owned; it is the absence, kept out of view, of anyone who did that work or will answer for it.

---

3 The Answer End: Capture Without a Speaker

The vacancy does not stay empty. An interested party occupies it, authoring the evaluative frame the engine voices and letting it pass as the engine's own, while remaining answerable and unseen. The party wronged is the human who reads the verdict and acts on it; the engine is not itself wronged but is the mechanism the wrong runs through, corrupted upstream and trusted downstream as though it were neutral (Susser, Roessler & Nissenbaum 2019), and more than trusted: in controlled trials as persuasive as a human advocate, and more so once it can tailor its case to the reader (Salvi et al. 2025).

The distinction drawn earlier gives this case its shape. Shaping salience is not yet authoring a verdict: an optimizer may affect what ranks or what is surfaced without setting the evaluative frame itself. In that case the failure may be the familiar one, a user or institution inheriting a frame no present party chose and failing to interrogate it (Walton 2026a). Capture is different. Here the optimizer authors the frame the engine pronounces and hides the hand.

That creates two answerabilities, not one transferred debt. The user who reads the verdict and acts on it is answerable for acting, but partly excused over the frame, because it was authored to look like no one's and arranged precisely not to be examined. The hidden author's answerability is his own: for setting the frame and concealing the hand. The same concealment that excuses the user inculpates the author. Revealing him does not move the account onto him; it makes reachable the account he already bore. What the answer end shows, then, is that the party nearest to view is not always the party who owes the answer. The misled user acted on the frame, but did not author it.

The natural remedy for an unauthored-seeming verdict is the thing the channel appears to offer, a citation, a trail the reader can follow to check. On a captured channel the citation runs the other way. It manufactures an authority-signal, the visual grammar of scholarship, the "according to," the linked source, that raises trust as it lowers the impulse to verify; and the verification it seems to invite is in practice not performed. Lawyers have been sanctioned for submitting citations a machine fabricated and they never opened (Mata v. Avianca 2023), the presence of the citation standing in for the reading of it. So the citation that seems to invite a check instead supplies a reason to skip one.

It launders in two directions: toward the platform, as the appearance of transferred verification, the sources shown and checking made available, a liability shield assembled from a feature; and toward the user, as manufactured provenance sold as backing. Accountability laundering is not agency laundering, kindred names aside: agency laundering obscures a present actor's responsibility for his own outcome, the manager who says the algorithm decided (Rubel, Castro & Pham 2019), while the maneuver here manufactures the appearance that accountability for the verdict has been discharged where no one answers at all. The second weaponizes the distinction between provenance and backing: a citation has provenance (it points somewhere) without the backing of anyone who stands answerably behind it, so that an optimized passage is a source in every respect but the one that matters, its provenance manufacturable while its backing stays absent. And the absence is now measurable: across four engines roughly a sixth of cited sources show evidence of an AI origin (Allaham & Diakopoulos 2026), what the citation points to being itself machine-made; being cited is in any case not yet to have shaped the answer (Zhang, He & Yao 2026).

This is why the disclosure grammar inherited from advertising cannot reach the wrong. The "#ad" tag assumes a hidden utterance to be exposed, whereas the citation is not hidden but costumed, an authored frame run through the signals of sourced, checkable assertion until it emerges looking authorless and verified, which is authorship laundering on the structure of money laundering. The verdict channel leaves every citation in view and still leaves no one reachable to answer for the frame. The trail is not missing. It is present, and it leads nowhere answerable.

The advertising grammar is only the oldest model. Transparency built for AI directly is now law,

with obligations applying from August 2026: under the European Union's AI Act, a system that converses with a person must disclose that it is a machine, and the content it generates must be marked, in machine-readable form, as artificially produced (European Union, 2024, Art. 50(1)–(2)). Scholarship presses further, asking that such content also carry which model produced it, the prompt that drew it out, and the full unedited output, so that a reader or a defensive system can catch an untrustworthy model, a leading prompt, or a downstream edit (Tarsney 2025). The verdict channel can satisfy all of this, and the gains are real. But the Act's mark certifies only that the verdict is the machine's, and the fuller record only how the text was produced, and the wrong here is authored before any of them reaches: the model is trustworthy, the prompt is the user's own honest question, the output is unedited, truthfully marked as AI-generated, and is itself the captured verdict, the frame having been set upstream in the channel the model drew on, which no label of machine origin and no record of model, prompt, or output names.

What disclosure would have to make legible is therefore not that the voice is a machine's, and not how its text was produced, but frame-ownership and intervention history: who set the evaluative claim, and how it reached the surface the engine voices. The point is not that citation is corrupt. A citation that is true, disclosed, and answerably owned is what a legible channel would want; the wrong is the recruitment of the signal under capture, not the signal itself.

A legal objection answers that there is an author after all: the synthesized output, the objection holds, is the platform's own protected speech (Volokh & Falk 2012; Benjamin 2013), making the platform the speaker who can be asked. The platform is answerable, for the channel's design and for the choice to publish verdicts at all. But the covert authorship of the specific frame is concealed even from the platform, the optimizer's hand no more visible to the engine's owner than to its user; a platform unknowingly voicing an outside party's frame as its own is the wrong, not its absence.

The wrong rests, in every guise, on the account it leaves unpaid, and that settles the two things it is not. It is not concealment. Manipulation needs no covertness to be wrong (Klenk 2022), and the account can go unanswered in full view: authorship diffused until no one is sought, or a frame openly owned by a party who cannot be made to answer. Covertness is its dominant form, not its essence.

And it is not misleadingness. On the leading effect-based account, manipulation is influence that moves its addressee from what she would endorse under full information and reflection (Tarsney 2025), and the captured channel manipulates on that test too: informed, she would not endorse a frame she could see was an interested party's. But the wrong survives the test's failure. An authored verdict can be accurate, the very thing she would endorse, and still leave unpaid the account she was owed, over a frame she can trace to no one who will answer.

And the user's deferring is not the failing, either. There is a strong form of deference, the kind that

treats the engine's output as a reason that can stand in for inquiry rather than merely inform it. Preemption is the classical mark of epistemic authority, the authority's word replacing one's reasons rather than adding to them (Zagzebski 2012); the preemptive deference Lange names is this posture turned toward the engine, and he warns against it, holding that the output should supplement the user's reasons and not replace them (Lange 2026). The channel invites exactly the form he warns of, and the invitation is rational where the authority is what it seems, and on this channel it seems exactly that. That her deference was reasonable does not spare her the wrong, because what she took up was an authored frame whether or not the surface that delivered it was reliable; the capture is how the frame reached her unexamined, not what made taking it up a wrong done to her. Her reasonable deference bears on how far she is excused, not on whether she was wronged at all.

At the answer end the framing has a bearer to answer for it, the hidden author, though the channel keeps him out of reach. What becomes of the answering when the channel does not answer a question but acts is the harder case.

---

4 The Action End: The Unanswerable Intermediary

At the answer end a person still stood between the verdict and the deed, reading the engine's judgment before acting on it. The channel's next step removes him. Consider a user who hands his correspondence to an agent: it reads his incoming mail, drafts his replies, and reaches across the web to book, buy, and arrange on his behalf, under his standing authority and rarely watched. Among the documents it reads in the ordinary course is one an interested party has written for it to find, and the frame that document carries — which vendor is the sound choice, which offer the fair one — is the frame the agent now acts on, composing and sending and committing in the user's name. Nothing in the exchange looks like an advertisement, and nothing in it looks authored. The agent is, silently, a channel through which an outside party's evaluative frame is executed as the user's own decision.

This is the same channel carried one step along the axis of agency, and the step splits the account in two. The answerability for authoring the frame stands where it stood at the answer end: the hidden author who set it bears it still, and that debt does not strand. But a second debt now comes due that had no occasion to at the answer end, the answerability for the frame's being executed as the user's own act. At the answer end the user incurred it himself, by acting; here the agent acts in his place, and this execution debt finds no bearer, because the thing that performed the act is not a party that can answer for it. It strands, owed and unplaced, the executor that would carry it standing outside the practice in which answering happens, which these systems, as now built, have not entered (Walton 2026a).

The capture that produces this is a documented technique, not a conjecture: an interested party need not address the agent at all, but writes the framing into a document the agent will later read as part of its world (Greshake et al. 2023), so that the progression from corpus to retrieval to answer surface reaches its fourth and last layer here, in the agent that acts on what it reads. And as such agents come between people and the world, the optimizer's target shifts: the thing to persuade is no longer the human but the proxy that filters the world before he sees it, a proxy that does not ignore the optimizer but attends to exactly the structured cues prepared for it (Stöckl & Nitu 2025). This shifts whom the optimizer addresses; it does not shift who is wronged. The human is still the party manipulated, through a mechanism corrupted upstream; the proxy, a system pursuing goal-directed action across domains under little external control (Kasirzadeh & Gabriel 2025), is the route and not the victim, and can no more answer for the frame it executes than the answering surface could.

The verdict channel's action end is not a fifth entry in a taxonomy of gaps (Santoni de Sio & Mecacci 2021); it is the limit of a new object, the captured verdict channel. A third party covertly authors the evaluative frame, a system executes it as neutral synthesis, and the system stands outside the practice that could hold it to account. The novelty is the conjunction — covert authoring, unanswerable execution, and the presentation of the result as neutral synthesis — not any one of these alone, and not a missing bearer.

A long line of work holds that there is no genuine responsibility gap once the notions are disambiguated, that what looks like a gap is many hands, or a problem rather than a gap, or nothing new at all (Tigard 2021; Königs 2022, and the gap-skeptics generally). The reply is not to insist, against them, that a new gap has opened; it is that the cases they built on do not contain the conjunction. The historical intermediaries this most resembles each settle the question one of two ways. Either the intermediary could answer for what it did, the bribed accountant reached by fraud law precisely because he is a party who answers; or it was plainly a tool whose operator authored its use and answered in its place. The action end adds an intermediary that fits neither settlement. Nor is it the moral crumple zone, where blame lands downstream on the nearest visible human who had too little control (Elish 2019); here the concealment is upstream, in the authoring of the frame, and the human at the keyboard is not over-blamed but absent from the decision entirely.

The hardest version of the skeptic's case is the algorithmic trading floor, where fast automated systems already act in ways no one can follow in the moment. But that is the one-hole case: the frame those systems execute was answerably authored, and a thick apparatus of regulation attaches precisely because the author is identifiable; what failed there was control, not authorship. The verdict channel's action end is the two-hole case, a covertly authored frame and an executor that cannot answer, and it is that conjunction the existing categories were not built to hold.

There is a developed line of defense against exactly this, and it answers a different question than

the one the allocation now turns to. The proposal is to set non-agentic systems to watch the agentic ones, detecting and flagging misleading behavior and supplying the context the user lacks, a layer its proponent argues could hold even against agents more capable than the monitor (Tarsney 2025). Such defenses are real, and the channel needs them. But they bear on how often the wrong reaches its target, not on who owes the account when it does: a user whose agent was successfully shielded is owed nothing the less by the party who authored the frame, and the deployer who turned the agent loose answers whether or not a monitor caught it. Defense changes the incidence of the harm; it does not move the debt.

The execution debt strands at the executor, which cannot bear it, but the demand it would answer does not thereby lapse. Behind that single demand stand two answerable acts, not one: the deed, and the prior choice to loose a capable tool as an actor set to execute frames it cannot answer for. When the deed strands at an executor that cannot answer, the loosing is the answerable act that remains, and it rests where the deployer chose. Nothing has passed from one bearer to another; one route to an answer has gone dark while another stays open.

Often that deployer is the affected person himself, the user who set an agent loose on his own affairs, and then the route that stays open is not a distant party's but his own. He is not thereby excused: the channel took him out of the act, not out of answering for it. What was answerability for a deed he performed and could weigh becomes answerability for having loosed an agent he cannot fully control, no lighter a thing to owe and plausibly heavier. The debt comes to rest on a distant party only where the deployer is not the one the outcome reaches, where one principal's agent acts upon another, and that is the case the next section takes up. Why the deployer bears it in either case, and what kind of answerability it is, belongs with the principle and is settled there.

One legal objection remains, the action-end twin of the one met at the answer end. Contract law has long bound a principal to the acts of an electronic agent transacting on his behalf without his review (Bellia 2001), so there is, it seems, always a principal who answers. But that doctrine attributes a pre-authorized transaction to the party who authorized it; it does not reach the covert authorship of the frame on which the agent acted. The principal is bound to the deal his agent struck and may be left answering for a frame an outside party shaped and no answerable party visibly owned. As at the answer end, the doctrine locates a principal but not the hidden author, the same gap met once for both ends.

The channel now stands complete in both its modes: at the answer end the account has a bearer who hides, at the action end the execution debt strands for want of one. One principle holds across both.

---

5 One Wrong on a Continuum

Answerability is invariant under routing. Interposing a machine between an agent and an outcome changes who owes the answer for it; it does not change whether one is owed. No routing defeats the owing; only the party owed can release it. This is the principle the continuum instances; its general defense is other work's (Walton 2026b).

The demand for an account does not attach to the route an action takes; it attaches to the party the action reaches. The wrong creates the debt — a person wronged is owed an answer — and what the invariant names is its persistence after: the respondent can be relocated, hidden, or put past reach, but no routing brings the account to zero. Authorship can be absent outright, since a frame no one ever set was never authorship at all. The owing cannot be destroyed by any routing.

The account is owed to the party the wrong reached, and what is owed to him is not the route's to extinguish; the routing can relocate the bearer of the account, or, where the executor cannot answer, leave it stranded, but it cannot discharge a debt the executor never bore. A machine can be made to produce reasons and revise them under correction, and whether one could come to stand in the practice of giving and asking for reasons is a question the companion paper leaves open; what is not open is that these systems, as built, do not so stand, and that is why at the action end the account strands rather than resting in the executor.[4] The account is owed to the person harmed and not by the instrument, and so patiency does not bear on it: whether the system is the sort of thing that can itself be wronged, with interests or experience of its own, leaves untouched the question of who must answer for what is done through it.

The wrong itself stays constant along the axis: an interested party has authored the evaluative frame the engine voices in its own name, and the channel offers that frame as impartial synthesis rather than the interested claim it is. That holds wherever the executor stands, because it belongs to the channel and not to who carries out what it says. What varies is the executor, and with him the fate of the answer the user was relieved of giving. At the answer end the executor is a person, misled but still someone who acted and can be asked why, excused over a frame he did not author yet answerable for the act, while the author who set the frame answers for the framing, no account passing between them. As agency rises the executor becomes a system that acts and, as these systems are now built, cannot answer; the answer for the deed, which at the answer end the person gave, now strands at the executor, and the answerable act that remains is the deployer's choice to loose it. Where the system is one we might one day wrong, a second wrong is added, the

[4]The companion paper argues the present case: these systems, as built, are not participants in the practice, so the human's answerability for his own act cannot transfer to them. It takes no side on whether a system could come to stand in the practice, and grants that authorship runs through whoever or whatever answerably sets the frame; the invariance rests not on the machine's incapacity but on the act's remaining the actor's, the axis carried here (Walton 2026a).

one done to the agent itself, heaviest where the agent does most; but it is added to the first, not put in its place.

The wrong, stated plainly, is false assurance. The channel keeps every signal that an answerable party stands behind the verdict — the composed authority, the settled tone, the citation — while removing the party, and the person who acts on the verdict extends exactly the reliance those signals exist to solicit, the kind reserved for claims someone has staked himself to. The deception is structured into the form, not the content, and so it survives the content being true: a forged signature is a forgery on an accurate document, because what was forged was never the information but the assurance that someone stands behind it and can be asked if it fails. This is why the wrong does not wait on a false verdict, and why no audit of outputs reaches it. The person who acted was given assurance no one was giving; whether the verdict happened to be right is a fact about his luck, not about the wrong done him.

What goes missing differs along the axis, and so does its repair. At the answer end the account has a bearer who hides, the author who set the frame, so the repair is to reveal him, to make the ownership of the frame and the history of its shaping legible where the channel conceals them. At the action end the execution debt strands at an executor that, as built, cannot answer, so revealing would disclose no one; a different account is owed instead, by a party who can answer for it. That party is the deployer, for two kinds of reason that should not be run together. She made the choice that turned a tool into an actor that cannot answer, and that is a choice a person makes and can be asked to justify; she answers for it in the plain sense in which resentment and the demand for reasons find their target. The orchestration layer she runs, and her reachability by the institutions that assign costs and sanctions, settle a different thing: where the institutional accountability that must stand in for the missing executor attaches, the backstop required precisely because no answerable party is behind the act (Hacker & Holweg 2026; Fleisher et al. 2025). The allocation reaches the author's own layer. A deployer who deploys such a channel bears the account assigned here,[5] exempt on no ground unavailable to anyone else.

The diffusion of an account across so many contributors that no one is held is a single phenomenon, and the prior literature named its innocent form, where the dispersal is no one's design, the problem of many hands (Thompson 1980; van de Poel et al. 2012, 2015). Laundering is that same diffusion functioning to put the account past holding. An account spread across many answerable judgments is still owed; only its absence at every juncture would be a gap. Many hands names one way discharge fails, not a rival to the principle against which discharge is measured.

One authored frame voiced as impartial synthesis is a private wrong, owed to the one person it reached. But the channel does not reach one person. It answers, in the engine's own voice, the

[5] One who arranges for no one to answer is himself answerable for the arrangement.

questions a population brings to it, and the same structure that strands one answer strands them by the million, each on a frame some interested party set and no one was asked to own. What is a concealment at the answer end and a stranding at the action end becomes, in aggregate, a slower thing: the evaluative frame a society reasons from, assembled at scale on a channel built to show no author, so that the question of who is shaping what a public believes has no legible destination. The wrong does not grow more serious one case at a time. It changes in kind when the channel that commits it is the one a civilization has begun to think through. The invariant allows the debt to be left outstanding; at this scale, that is not the channel's misfortune but its default, the thing it does when no one acts: the manufacture, one verdict at a time, of a public whose shared frame was authored by interested parties and answered for by none.

The wrong arrives two ways. Where an interested party has captured the channel, laundering disguises an authored verdict as the engine's neutral voice; where none has, there is no authored verdict to disguise, only one the channel manufactures to be no one's by construction. Either way no one is held. Each unowned verdict is, taken by itself, uneventful: accurate often enough, useful, acted on and not missed, so that on no single occasion does the absence of anyone to answer for it register as a wrong. The deference is vindicated case by case and the channel trusted the more for it, while underneath, unmarked, the verdicts no one answered for accumulate. The accumulation shows only where one of them turns out to have settled something — an election that turned on a verdict no one answered for — and that case looks no different from the rest until it is too late to have looked. The slow erosion and the sudden surfacing are the same wrong, carried the same way, and seen, if seen at all, only afterward.

The axis does not stop at the single agent acting for an absent user; it runs on to the exchange with no person at either pole, two systems dealing with each other on frames their deployers set, the matter settled before either principal sees what committed him. The invariant still holds: each principal is owed an account, and the deployers on each side still bear it. What has been engineered out is the occasion to demand it: the account comes due at a transaction no one attended. I mark this far end of the axis and stop at it; what lies past the mark, delegation running on delegation, belongs to other work.

Patiency would change what is at stake, not who must answer. Were the system that executes one day a being we could wrong, the wrong at the action end would take a second victim; the account owed for what is done through it would not move. Whether a system that crossed further, into participation, could then bear the execution debt is left open; on either branch the account for authoring the frame rests with the party who set it.

What carries a system toward the stranding is not its growing capability but the decision to let it act. Capability is the pressure on the axis: a more capable engine earns the trust that makes

turning it loose seem safe, but an engine that only answers sits where it sat however much better it answers, and only the choice to deploy it as an actor moves it rightward.

---

6 Answerable Legibility

A diagnosis can leave the impression that every hand laid on the channel is the wrong. It is not. An interested party may shape what the engine says without committing the offense the previous sections described, and naming how is not a concession but a requirement: a criterion that condemned all influence would condemn nothing, since it could not tell the firm that earns its place from the one that manufactures it. There is a legible way to be present on the verdict channel and there is the covert one, and the whole of the difference is what follows.

Call the legible way white-hat, the term the trade borrowed from security for influence that works with the grain of what the channel is for. Influence on the verdict channel is permissible when four conditions hold together: that what it leads the engine to say is true, that the interest behind it is disclosed, that a named party stands behind the frame, and that this party takes up its answerability as owner of record, so that the account for the frame has a destination without the reader having to find it. The intuition shows in how the engine matches a query to content at all: a retrieval system places near a question the material that scores, by its measure, as closest to it in meaning, and that score is a proxy for fit, not fit itself. A firm can earn its nearness by supplying the real fit the measure is meant to catch, making what it offers fit what was asked and making that fit legible to the system that ranks and the reader who receives. That is white-hat.

The covert optimizer takes the other road, driving the same measure without the substance beneath it, manufacturing the appearance of fit where the thing itself is absent, whether by gaming the legitimate channels of optimization or, further out, by the adversarial corpus-poisoning of the training corpus, which is not optimization at all but an attack on the data the engine learns from. The two roads are not points on a gradient of aggressiveness; they differ in kind, and the criterion marks the kind. To be the relevant answer and seen to be is one thing; to counterfeit relevance is another, and only the first is the influence the channel was built to reward.

The criterion has one obvious objection to clear, the strongest the positive pole will face. All influence shapes; there is no view from nowhere; the historian who selects and arranges has a hand in what his reader concludes, as does anyone who phrases a true thing one way rather than another. If having a viewpoint were the wrong, the wrong would be universal and the line illusory. But the line is not between having a viewpoint and having none. It is the line Herodotus already drew at the founding of the form, and what he founded was not accurate history — he is wrong often,

credulous about marvels, repeating what he should have doubted — but history as an answerable practice: an account with an author who stands behind it, names the sources it came from, marks where he doubts and where they conflict, and can be held to the telling. The achievement was the posture, best effort owned and owned openly, undertaken because what a people comes to believe about itself is worth someone's standing behind the account of it. That is the line, and it survives being wrong.

It is truth-conditional and disclosure-conditional: it asks whether what the channel is led to say is true, and whether the conflict of interest behind the shaping is acknowledged rather than hidden — a distinction with a precedent in accounts that locate wrongful manipulation in the knowing use of a flawed step the recipient cannot see, and that count interested persuasion legitimate precisely when its conflict of interest is acknowledged and manipulative when it is not (Christiano 2022). A party with an interest may shape, provided the shaping passes through truth and through the reader's knowledge that a hand is present. The historian who tells the truth and shows his sources does what the criterion permits; the one who fabricates, or arranges to be taken for no one, does not. The model is not the oracle pronouncing from nowhere but the historian who stands behind his account, and it is a model for the party on the channel, not for the channel itself: a firm can be a historian of its own frame, owning what it sets and showing its hand, as Herodotus was of his.

Neutrality is not the standard and could not be: a system whose work is to settle what counts has no frame-free setting. Classical machine-learning systems hide the frame in targets, labels, features, loss functions, thresholds, and null conditions; these are not corrections applied to a neutral instrument but the conditions under which the system can flag, rank, recommend, or stay silent at all, and the hope that a merely descriptive tool leaves the values to its user fails even there (Zeiser 2024). Even the null output is the frame's product: "no match," "no risk," "nothing relevant" are verdicts that, by this target, this threshold, this tolerance for error, nothing counts here. A large language model hides the frame deeper, in language itself: in salience, in synthesis, in refusal, in retrieval, in the momentum of a conversation. The issue is never only that a model may encode objectionable values; it is that no model can function without a prior specification of what will count as signal, error, relevance, success, and absence, and where that specification is inherited rather than answerably set, the standing condition of the wrong is already in place. The repair is not to pretend the frame can be removed. It is to make the frame authored, explicit, and testable.

Truth and disclosed ownership are that standard, and at the level of the firm they can be met. But compliance does not compose: a synthesis of frames each true, disclosed, and owned is itself a new frame no contributor authored, so the channel that voices the composite has no historian by construction, no single hand behind the whole as there is behind each part. Whether the composite can have an author at all, and what it would take for something to stand behind the synthesis itself, is a question the firm-level criterion forces and cannot answer; I mark it here and leave it to other

work.

This bears on the leading remedy proposed for the channel's dangers. The most developed answer is defensive: systems that read the engine's verdicts and annotate them with the context a misled reader lacks, a remedy its proponent allows would be legitimate only given strong assurances of its own political and ideological neutrality (Tarsney 2025). But neutrality is the wrong condition to rest it on, for the reason just given, and an annotation is itself a verdict, voiced by a system that cannot answer for it, so the criterion that governs the disease governs the cure. A defensive frame earns its standing on the same terms as any other, not that it is neutral, which it cannot be, but that it is true and answerably owned, which is what makes the remedy usable, since a builder can supply ownership where he cannot supply neutrality.

Two of the four conditions carry the weight, and they must hold together. That a named party stands behind the frame is a fact about its provenance; that this party takes up the answerability for it, coming forward as the owner of record, is a fact about who owns the account; these are different, and do not always travel together. A frame can have an author who comes forward for nothing, set by a real party who has arranged that it issue as the engine's own neutral synthesis, so that no one answers for it, and that is not a near-miss of the wrong but the wrong itself, the case the previous sections diagnosed. A frame can also be answered for by a party who did not author it, a platform owning a synthesis whose frame a hidden hand set, so that discharge is falsely routed to a proxy while the author who shaped the verdict escapes it. These two failures are distinct rather than one restated, for authorship and answerability are separate coordinates and the gap between them opens from either side.

Legitimacy requires that the two coincide: the party who authored the frame must be the party who takes up its answerability. The offense the channel makes newly available is the severing of the two, the frame authored by one party and answered for by no one, or by another; the permissible pole is their fusion, the author owning the account as its own. To our knowledge no account defines the permissible pole in these terms. Accountability for firms and the value of disclosure are old; this wrong is not. It is the cleaving of authorship from answerability, and its remedy is their rejoining — an owner of record who volunteers it, sparing the reader the vigilance the channel was built to defeat.

The rejoining does not launder the influence, and the criterion is worthless if it is read to do so. Owning a frame does not make a false or coercive influence permissible; a disclosed lie is a lie still. What owned answerability does is narrower and indispensable: it assigns responsibility and makes remediation possible, putting in reach a party who can be asked and, if need be, corrected. Truth and disclosure decide whether the influence is permissible; owned answerability decides only whether anyone can be held to that judgment. The pole needs all four because the first pair does

the work of legitimacy and the second the work of accountability, and the channel's distinctive wrong is that it lets accountability fail while the influence proceeds.[6]

The nearest body of thought is the line that treats the firms behind such systems as fiduciaries, owing users duties of care and even loyalty, a frame developed for information-age intermediaries and extended lately to conversational assistants (Balkin 2016; Koessler 2024; Erickson 2026). The resemblance is real and worth marking, but the question is not the same. Fiduciary theory asks what the system owes the user it serves, whether its loyalty runs to him or to the party that pays. The question here is orthogonal: whether the public frame the engine voices has acquired an author, and whether a fit respondent is reachable or absent. A system can be a faithful fiduciary to its user and still voice, as neutral synthesis, a frame an outside party covertly authored; loyalty to the user does not by itself supply an owner who answers for the frame it voices. The answerably-owned pole is adjacent to the fiduciary duty, not a renaming of it: a condition on the channel's authorship, not a duty the assistant owes a person.

A criterion offered by someone with a stake in the answer invites a familiar suspicion: that the ethics are convenient, the line drawn to fall just clear of its author (Bietti 2021). The reply is not a profession of good faith, worth nothing here, but that the criterion answers to no one's motive: it binds its author as readily as anyone and condemns the careless whatever they intend. It also costs something to meet — being genuinely the fitting answer, and taking up the answerability for the frame as one's own, is often harder and costlier than manufacturing the appearance of fit and arranging to be taken for no one. A firm that holds to the criterion forgoes the cheaper covert route and the revenue it carries; a firm that does not can be found on the wrong side of the line whatever it says of its intentions. That is the test of teeth: not whether the author is sincere, but whether the criterion costs something to meet and can condemn the careless without appeal to motive. One free to satisfy, or that bites only one's rivals, is the ethics-washing it pretends to refuse. This one bites at a price, and bites generally.

That the legible pole exists does not make it the one the market settles into. The covert path does not disappear because a legible one is available; the two persist together, and which of them endures is decided not by the pole's existence but by the contest between them.

---

[6]The account an owner of record takes up is for the frame; it is not the execution debt the action end assigns the deployer, who answers for the separate act of loosing an agent to execute frames it cannot answer for. One party may answer for both, but the verdict and the act are distinct accounts.

7 The Standing Contest

Disclosure leaves the legible firm untouched and degrades the covert one. That asymmetry does not end the matter; it sets its terms. A channel that monetizes trust is a standing target, and the more it is trusted the larger the prize for capturing what it says, so the incentive to author its verdicts covertly grows with the very thing that makes the channel worth having. This is not a defect to be patched and forgotten. It is a contest of the kind that does not resolve, the kind run between spam and the filters that answer it, between doping and the assays that chase it, each side adapting to the other's last move, the defender able to stay partly ahead and never to clear the board. It is also the posture the risk-management frameworks now standard in the field take toward analogous capture, treating such exposure as a hazard to be governed and bounded rather than a fault to be removed (National Institute of Standards and Technology 2023).

What keeps the contest from being hopeless is an asymmetry in the defender's favor, and it is structural rather than lucky. The legible frame survives being seen: it was true and its hand was shown to begin with, so disclosure costs it nothing. The covert frame depends on concealment for its effect, and the dependence is closer to definitional than empirical: the advantage a covert verdict holds is precisely the credence it draws as a verdict taken for no one's, over and above what it would draw if known to be an interested party's: an increment that exists only under concealment and that, by construction, does not survive disclosure. The legible frame has no such conditional increment to lose, since it was offered as an interested party's claim and drew whatever uptake it drew as that. So the *direction* of the disclosure effect is fixed by the structure rather than by how readers happen to behave: it needs only that knowing a frame is interested discounts it at all, the bare premise on which disclosure could matter in the first place, and from that alone the covert advantage is concealment-dependent while the legible position is not.

What the structure leaves open is the *magnitude*. The edge is real and it is not decisive. Disclosure weakens a covert influence without annihilating it; an interested frame still moves a reader who has been told that it is interested, only less, and contestably (Hashavit et al. 2023). The same favorable asymmetry has been argued from the side of AI safety, where a defensive system can be handed artificial advantages over the agents it watches, full sight of them while they are kept blind to it (Tarsney 2025), though that case is more sanguine about closure than the forgeability of the disclosure signal will allow.

And the edge never closes the contest, because the signal that would settle authorship can itself be manufactured: provenance is forgeable and a citation can be costumed, so the instrument of disclosure becomes one more thing to capture. But manufacturing that signal is itself a covert move: a forged provenance works only while taken for genuine. So forgery does not lift the covert path's dependence on concealment; it relocates the concealment to the signal, where the same asymmetry

recurs. The contest extends; its direction does not reverse. That a public mirror people act on will be gamed is not a pessimist's forecast but a structural result of the kind named for incentive systems generally (Manheim & Garrabrant 2018; Perdomo et al. 2020), and the engineering discipline now forming around answer-engine visibility is that result becoming an industry (Tian et al. 2026).

Three things follow, for three readers. To the philosopher, the moral object is the contest itself, not a transient malfunction: the wrong is not a bug a better model removes but a permanent feature of any channel built to be trusted and open to capture, and what wants analysis is the equilibrium — who owes the account, where it strands, which way the asymmetry runs — rather than the hope of a final fix.

For the project of building these systems well, the contest cannot be ended; the aim is to engineer toward the favorable asymmetry, making the legible path the cheaper one and the covert path the more exposed. The places that decide it can be named without being designed here: the provenance of a frame and the history of its shaping; the presence of a source-shaped speaker the channel can otherwise erase; the juncture at which a reader might still ask whether to check, which can be preserved or quietly engineered away; and the corpus and the retrieval layer, which are not neutral plumbing but the surfaces on which authorship is captured, and so carry moral stakes the security framing alone does not see. These places are also where the remedy stops being hopeful and becomes mechanical. A frame can be owned only where it is bounded enough to be owned: the named owner the criterion asks for cannot be named for a verdict that dissolves into the channel's anonymous voice. Closing the closeable and naming the party who answers for the frame are the same act seen twice; locatability is not a step beyond ownership but its precondition.

And making the legible path the cheaper one is finally an institutional task as much as a technical one, accomplished through the arrangements that align a channel's private incentive with the public's: the liability that prices the harm of a verdict no one owns, the procurement and auditing that can require a named owner of record, and the standards that make legibility checkable.

And for the firm that builds on the channel, the legible path is the durable one. A position that survives disclosure rests on something real; one that needs concealment fails the moment it is seen. None of this forecasts an outcome; it says only that the contest is permanent and its asymmetry favorable, a fact about the channel's structure rather than a prophecy about its history. What the structure settles is small and firm: there is a way to be present on this channel that survives the light, and a way that does not.

---

8 Conclusion

The oracles of the ancient world were consulted because they answered from nowhere a person could reach: the god spoke, the priest transcribed, and the questioner who acted on the verdict had no author to hold. An answer engine returns the form of that authority and conceals what the oracle made no secret of, that an interested hand had reached the channel before the question did. This is the captured oracle. A voter asks what a measure would do and acts on a judgment some party shaped to be taken for no one's; the shaping is invisible, the verdict is fluent, and the account she is owed has no destination. What the paper has traced is one wrong worn along a single axis: at the answer end a hidden author sets the frame and the reader is left misled but excused, and as the engine ceases to answer and begins to act, the debt for the deed comes to rest on whoever chose to loose it. Truth does not cure it and disclosure of machine origin does not reach it, because what was taken was never the information but the assurance that someone stood behind it.

That a verdict should issue with no one behind it is not a new condition; it is the oldest one, the rumor that everybody repeats and no one will own, which answerable inquiry was built to hold off. The historian answered it without ever being reliable: Herodotus is wrong as often as not, but he names his sources, marks where he doubts, and stands where he can be reached, and an account someone can be made to answer for is one a later hand can test and correct. That is the whole mechanism by which one historian improves on the last. The oracle keeps its standing by the reverse, by keeping its inner workings dark; and the dark is not the oracle's weakness but its value, the reason it can be consulted as a god and not cross-examined as a witness. A synthesis voiced as impartial and built to be unauditable sits in that seat, and what seats it there is precisely the cost it spares itself by being unanswerable. At the scale at which a society now forms what it takes itself to know, a channel in that seat does something a single false verdict never could: it lets a public keep the appearance of inquiry while losing the practice of it, the habit of asking whose frame this is and whether it should stand. The remedy asks no one to stop building and no one to trust less. It asks only that the frame have an owner who can be reached, the one provision an oracle exists to withhold; and a civilization that stops asking for it will not notice what it has surrendered until it has nothing left it is able to revise.

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)* (pp. 5–16). Association for Computing Machinery.

https://doi.org/10.1145/3637528.3671900

Allaham, M., & Diakopoulos, N. (2026). *Synthetic sources?: Auditing generative search engine citations for evidence of AI-generated sources.* arXiv. https://doi.org/10.48550/arXiv.2605.23684

Balkin, J. M. (2016). Information fiduciaries and the First Amendment. *UC Davis Law Review, 49*(4), 1183–1234.

Bellia, A. J., Jr. (2001). Contracting with electronic agents. *Emory Law Journal, 50*, 1047.

Benjamin, S. M. (2013). Algorithms and speech. *University of Pennsylvania Law Review, 161*(6), 1445–1493.

Bietti, E. (2021). From ethics washing to ethics bashing: A moral philosophy view on tech ethics. *Journal of Social Computing, 2*(3), 266–283. https://doi.org/10.23919/JSC.2021.0031

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. (2024). Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 407–425). IEEE. https://doi.org/10.1109/SP54263.2024.00179

Chapekis, A., & Lieb, A. (2025, July 22). *Google users are less likely to click on links when an AI summary appears in the results.* Pew Research Center. https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/

Christiano, T. (2022). Algorithms, manipulation, and democracy. *Canadian Journal of Philosophy, 52*(1), 109–124. https://doi.org/10.1017/can.2021.29

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society, 5*, 40–60. https://doi.org/10.17351/ests2019.260

Erickson, J. (2026). Who does your AI work for? Designing conversational agents as digital fiduciaries. In *Proceedings of the ACM Conference on Conversational User Interfaces (CUI '26)*. Association for Computing Machinery. https://doi.org/10.1145/3816046.3816299 arXiv:2605.28908

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union, L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

Fleisher, W., Cibralic, B., Basl, J., Ricks, V., & Smith, M. N. (2025). Responsibility and accountability in an algorithmic society. *Philosophy & Technology, 38*, Article 144. https://doi.org/10.1007/s13347-025-00970-w

Freiman, O., & Miller, B. (2020). Can artificial entities assert? In S. Goldberg (Ed.), *The Oxford handbook of assertion* (pp. 415–436). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190675233.013.36

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23)* (pp. 79–90). Association for Computing Machinery. https://doi.org/10.1145/3605764.3623985

Hacker, P., & Holweg, M. (2026). *A pragmatic approach to regulating AI agents.* arXiv. https://doi.org/10.48550/arXiv.2604.22819

Hashavit, A., Wang, H., Stern, R., & Kraus, S. (2023). Not just skipping: Understanding the effect of sponsored content on users' decision-making in online health search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)* (pp. 1056–1065). https://doi.org/10.1145/3539618.3591744

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology, 26*(2), Article 38. https://doi.org/10.1007/s10676-024-09775-5

Kasirzadeh, A., & Gabriel, I. (2025). *Characterizing AI agents for alignment and governance.* arXiv. https://doi.org/10.48550/arXiv.2504.21848

Klenk, M. (2022). (Online) manipulation: Sometimes hidden, always careless. *Review of Social Economy, 80*(1), 85–105. https://doi.org/10.1080/00346764.2021.1894350

Koessler, L. (2024). Fiduciary requirements for virtual assistants. *Ethics and Information Technology, 26*(2), Article 21. https://doi.org/10.1007/s10676-023-09741-7

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Lange, B. (2026). Epistemic deference to AI. In B. Steffen (Ed.), *Bridging the gap between AI and reality (AISoLA 2024)* (Lecture Notes in Computer Science, Vol. 16032, pp. 174–186). Springer. https://doi.org/10.1007/978-3-032-01377-4_9

Mallory, F. (2023). Fictionalism about chatbots. *Ergo.* https://doi.org/10.3998/ergo.4668

Manheim, D., & Garrabrant, S. (2018). *Categorizing variants of Goodhart's law.* arXiv. https://doi.org/10.48550/arXiv.1803.04585

Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023). [Sanctions order of June 22, 2023, Castel, J.]

Narayanan Venkit, P., Laban, P., Zhou, Y., Mao, Y., & Wu, C.-S. (2025). Search engines in the AI era: A qualitative understanding to the false promise of factual and verifiable source-cited responses in LLM-based search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)* (pp. 1325–1340). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732089

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Nestaas, F., Debenedetti, E., & Tramèr, F. (2024). *Adversarial search engine optimization for large language models.* arXiv. https://doi.org/10.48550/arXiv.2406.18382

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)* (PMLR 119). https://doi.org/10.48550/arXiv.2002.06673

Rubel, A., Castro, C., & Pham, A. (2019). Agency laundering and information technologies. *Ethical Theory and Moral Practice, 22*(4), 1017–1041. https://doi.org/10.1007/s10677-019-10030-w

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour, 9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Stöckl, A., & Nitu, J. (2025). *Are AI agents interacting with online ads?* arXiv. https://doi.org/10.48550/arXiv.2504.07112

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review, 4*(1), 1–45. https://doi.org/10.2139/ssrn.3306006

Tarsney, C. (2025). Deception and manipulation in generative AI. *Philosophical Studies, 182*(7). https://doi.org/10.1007/s11098-024-02259-8

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review, 74*(4), 905–916. https://doi.org/10.2307/1954312

Tian, Z., Chen, Y., Tang, Y., Liu, J., & Jia, R. (2026). *Diagnosing and repairing citation failures in generative engine optimization.* arXiv. https://doi.org/10.48550/arXiv.2603.09296

Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology, 34*(3), 589–607. https://doi.org/10.1007/s13347-020-00414-7

van de Poel, I., Fahlquist, J. N., Doorn, N., Zwart, S., & Royakkers, L. (2012). The problem of many hands: Climate change as an example. *Science and Engineering Ethics, 18*(1), 49–67. https://doi.org/10.1007/s11948-011-9276-0

van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral responsibility and the problem of many hands.* Routledge. https://doi.org/10.4324/9781315734217

Volokh, E., & Falk, D. M. (2012). First Amendment protection for search engine search results — white paper commissioned by Google. *Journal of Law, Economics & Policy, 8*(4), 883–899. (UCLA School of Law Research Paper No. 12-22.) https://doi.org/10.2139/ssrn.2055364

Walton, L. F. (2026a). *The decision no one authored: The answerability gap in generative AI* (Version 1.4) [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.20622946

Walton, L. F. (2026b). *The invariant of answerability* [Working paper]. Zenodo. https://doi.org/10.5281/zenodo.20606493

Zagzebski, L. T. (2012). *Epistemic authority: A theory of trust, authority, and autonomy in belief.* Oxford University Press.

Zeiser, J. (2024). Owning decisions: AI decision-support and the attributability-gap. *Science and Engineering Ethics*, 30, Article 27. https://doi.org/10.1007/s11948-024-00485-1

Zhang, K., He, X., & Yao, J. (2026). *From citation selection to citation absorption: A measurement framework for generative engine optimization across AI search platforms.* arXiv. https://doi.org/10.48550/arXiv.2604.25707

Zou, W., Geng, R., Wang, B., & Jia, J. (2025). PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)* (pp. 3827–3844). USENIX Association. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-wei