ACCEPTED MANUSCRIPT

The Captured Oracle

Authorship and Agency in the Ethics of Answer-Engine Optimization

Luke F. Walton
Independent Researcher, San Diego, California
ORCID 0009-0005-9263-1954 · luke@lukefwalton.com

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s43681-026-01532-2

Cite the Version of Record

Walton, L.F. The captured oracle: authorship and agency in the ethics of answer-engine optimization. *AI Ethics* **6**, 665 (2026). https://doi.org/10.1007/s43681-026-01532-2

About this file

The pages that follow are the manuscript accepted by *AI and Ethics* on 9 November 2026, converted to PDF without edits from the file submitted for double-anonymous review. That is why it opens "Author information removed for double-anonymous review" and carries no declarations: they were supplied separately and are reproduced below as published. The text has not been copy-edited or typeset.

This file is not distributed under a Creative Commons or other open licence. It is hosted on the author's own website, lukefwalton.com, under Springer Nature's accepted manuscript terms of use: https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

Declarations, as published in the Version of Record

Competing interests. The author is the founder and chief executive officer of Surmado, Inc., a company that builds and operates answer-engine and AI-answerability infrastructure related to the subject of this paper. The author is also a named inventor on a pending provisional patent application assigned to Surmado, Inc. and related to managed AI system architecture. The research received no external funding and expresses only the author's views.

Use of generative AI. Several frontier foundation models (from Anthropic, OpenAI, Google, and xAI) 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.

Ethics approval. Not applicable. **Consent to participate.** Not applicable.

The Captured Oracle

Authorship and Agency in the Ethics of Answer-Engine Optimization

Author information removed for double-anonymous review.

Abstract. Answer-engine optimization is SEO for large language models. Ask an answer engine a question with a contested answer and it may return a verdict — composed in the engine's own voice, sources nested in the answer — with no visible author. An optimizer can constitute the frame the engine voices as its own. A public acts on these verdicts with no author reachable: that is the wrong at the answer end. That the engine is only a tool leaves the case standing: where a verdict has been captured, the frame it voices is someone's, whom the channel conceals, leaving the user the only party in view. An interested frame voiced as impartial synthesis is the wrong's sharpest form: it survives accuracy and disclosure. Given such a frame, what varies is one thing: the agency of whoever executes it. At the answer end a person acts on a verdict whose frame was hidden; the hidden author must answer. At the action end an AI agent acts in their place: the executor cannot answer, and what remains 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.

Keywords: answer-engine optimization; answerability; authorship; AI agents; online manipulation; accountability

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. 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 that it can return a finished evaluative judgment, 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.

A prior failure lies on the user's side: a person may treat a system's output as evidence rather than a proposal and ratify a frame she never examined. The failure studied here lies on the other side of the same channel. Once an unauthored-seeming verdict is trusted, the verdict itself becomes the target. The optimizer is not merely trying to be found; he is trying to author the frame the engine will voice — the specification of what the answer is to count as — while preserving the appearance that the frame belongs to no one. Where the channel a reader defers to has been gamed this way, her deference launders his authorship as objectivity. The subject has moved from the one who defers to the one who optimizes, and the optimizer is the party 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). Optimization is a service: shaping what the engine says 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. The incentive this builds is easily illustrated: a national bank could out-optimize a regional one and have the engine return the larger budget as the disinterested recommendation, with no reader the wiser.

The closed and the live cases differ by one architectural fact: how outside material is admitted to the evaluative frame the system returns. On a closed system outside material is admitted at answer time only by the user's own hand, selected and supplied by her, though a document she selects can carry a criterion she never saw; the route that does not pass through her hand at all is to poison the training corpus, which is practical (Carlini et al. 2024) but weak against a frame already set; a recruiter whose screening tool inherits its definition of merit from a corpus she never examined works on this substrate. On a live-inference system the frame is assembled at answer time, from a channel anyone can write to, and the sharpest interventions are native to it — a handful of crafted passages in what the engine retrieves, sufficient in demonstrated cases to settle its conclusion (Zou et al. 2025), or an answer surface steered by content authored for the engine to ingest (Nestaas et al. 2024). The line is therefore structural rather than a verdict on which products are dangerous today, and it predicts its own crossings: let the recruiter's closed tool go agentic and reach the live channel, and it moves from shaping what is salient to constituting the frame that settles the conclusion. Substrate decides only how the material is admitted; once admitted, what varies is agency — who executes the verdict — and that axis is the continuum that follows.

The case the paper opened on is the sharpest available, because what is being said about a live measure is being said now, after any training cutoff: an engine

that answers it from current coverage is reading the same public channel interested parties can seed.

The hidden author already bears the account for the verdict that settles her question. Yet the channel leaves no one reachable for the settling. Hicks et al. (2024) name a neighboring failure, the engine's indifference to truth: that diagnosis concerns whether the output aims at being true. The wrong here is a settling no one can be reached for, and a true verdict can still leave her with it.

2 The Verdict Channel

What an answer engine now returns lies along a range. 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. The synthesized verdict issues from a frame. The frame, here and throughout, is the specification of what counts as signal, relevance, success, and absence in answering the question, including the weighting that settles which answer issues. The conclusion is what the frame issues.

The featured snippet, which began appearing above the ordinary Google search results around 2014, sits close to the attributable end. It answers in language that is visibly someone else's: a sentence or two extracted from a page, set apart and credited, with the page itself a click away, so that 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: an AI overview composes the answer in its own voice and nests sources against 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). The behavioral point is enough here: the engine makes the claim, defends it when pressed, and withdraws it when corrected. 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 nested in 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. Answerability, in the sense at issue, is the standing to be asked for the reasons and judgments behind a verdict, and to owe them, as distinct from whether an act is attributable to one's character or whether one is liable to sanction for it (Shoemaker 2011). Assertion carries a readiness to defend a claim; the answerability at issue 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. What has moved onto the answering surface is 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.

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, 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. Nothing in that arrangement is yet a wrong: it is, so far, a vacancy, empty and unwatched, the wrong's standing condition.

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 unseen. The party wronged is the human who reads the verdict and acts on it. The engine is the mechanism the wrong runs through: corrupted

upstream, and trusted downstream as though it were neutral (Susser, Roessler & Nissenbaum 2019). In controlled trials, answer engines are as persuasive as a human advocate, and more so once they can tailor the case to the reader (Salvi et al. 2025).

Shaping salience is not yet authoring a verdict. An optimizer may affect what ranks without setting the evaluative frame itself, and the failure there is the familiar one: a user inheriting a frame no present party chose and failing to interrogate it. Authorship is the intentional composition of an evaluative-frame element: a criterion, a weighting, a specification of what counts. Uptake is that element's actually entering the frame the engine assembles for the question; wording can be echoed and a passage cited without either determining the judgment. Capture is the condition in which the resulting frame is sufficiently constituted by the authored element to carry the interested judgment. The test is constitution rather than magnitude. An informal counterfactual marks the line. Remove the authored element and ask what the engine would then have judged by, rather than what it would have concluded. Where the standard of what counts survives the removal and only the phrasing or the citation changes, there was uptake; where the standard goes with the element, the frame was constituted by it. A verdict can turn on any retrieved source without that source having set the standard it turned on, which is why the counterfactual runs on the criterion. Causal contribution to retrieval, ranking, or phrasing is not authorship of the frame. Strategic intent whose element never enters the evaluative criterion has authored nothing the verdict issues from. A frame may have several authors, each of a part. Taken together, authorship composes an element, uptake admits it into the frame, and capture is the frame's being constituted by it; constitution is the test that tells capture from uptake.

Across forty-five reviewed studies, content already retrieved can causally alter its citation or use (Martinez 2026). That is the influence the constitution test already distinguished from authorship of the frame. No reviewed technique shows a stable, longitudinal, cross-platform effect on organic discoverability or downstream behavior, and this paper claims none. The object is the structurally available case in which an interested contribution becomes constitutive of the frame a verdict issues from, not a party holding the pipeline under commercial control.

Capture alone is still short of the wrong, since a captured frame that is answerably owned is the legible influence the four conditions permit; the wrong fires where the constituted frame is not answerably owned. Its commercially salient form is covert: the optimizer authors the frame the engine pronounces and hides the hand, and that concealment creates two answerabilities. 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; and 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. The party nearest to view is not always the party who owes the answer.

The natural remedy for an unauthored-seeming verdict is the citation the channel appears to offer. On a captured channel the citation runs the other way: it manufactures an authority-signal that raises trust as it lowers the impulse to verify, a substitution lawyers have been sanctioned for acting on (Mata v. Avianca 2023). The maneuver manufactures the appearance that accountability has been discharged where no one answers at all; the agency laundering already named in the literature, the manager who says the algorithm decided, obscures a present actor's responsibility for his own outcome (Rubel, Castro & Pham 2019). A citation points somewhere without the backing of anyone who stands answerably behind it. Advertising disclosure cannot reach this: the "#ad" tag assumes a hidden utterance, and the citation is not hidden but costumed.

Transparency built for AI directly goes further and is now law, and the law differentiates the obligations: a system interacting directly with a person must disclose that it is an AI system, unless this is obvious from the point of view of a reasonably well-informed, observant and circumspect person, given the circumstances and the context of use; providers must mark generated content, in machine-readable form, as artificially produced, except where the system performs an assistive function for standard editing or does not substantially alter the input; and deployers who publish generated text to inform the public on matters of public interest must disclose its artificial origin, exempt where the content has undergone human review or editorial control and a natural or legal person holds editorial responsibility for the publication (Regulation (EU) 2024/1689, Art. 50(1), 50(2), 50(4)). Scholarship presses further still, asking that such content also carry which model produced it, the prompt that drew it out, and the full unedited output (Tarsney 2025). The verdict channel can satisfy all of this, and the gains are real. But the mark certifies only that the verdict is the machine's, and the fuller record only how the text was produced, while the wrong here is authored before either reaches: the model is trustworthy, the prompt is the user's own honest question, and the unedited, truthfully marked output is itself the captured verdict, the frame having been set upstream in the channel the model drew on. Even the duties that exist certify machine origin and content class, not frame ownership; and Art. 50(4), the one provision keyed to a party's holding editorial responsibility, is the one that cannot reach a verdict voiced to a single user. What disclosure would have to make legible is therefore not that the voice is a machine's, nor 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.

Citation is not the corruption. 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. Two legal doctrines locate a speaker without reaching the author. The synthesized output may be the platform's own protected speech (Volokh & Falk 2012; Benjamin 2013), and contract law has long bound a principal to the acts of an electronic agent transacting on his behalf (Bellia 2001). Each doctrine finds a present party, the platform at one end of the axis and the principal at the other, while the authorship of the specific evaluative frame stays concealed even from that party.

The ground is unpaid answerability for the frame. What entitles the user to an answerable author is testimonial: an interface that addresses a person in the form of an answer solicits the reliance reserved for a judgment someone stands behind.

Manipulation needs no covertness to be wrong (Klenk 2022). The account can go unanswered in full view, whether diffused until no one is sought or openly owned by a party who cannot be made to answer. On the leading effect-based account, manipulation moves its addressee from what she would endorse under full information (Tarsney 2025). 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. The preemptive deference that lets an authority's word replace one's reasons rather than supplement them (Zagzebski 2012; Lange 2026) is rational where the authority is what it seems. On this channel it seems exactly that, so her reasonable deference bears on how far she is excused, not on whether she was wronged.

The framing has a bearer to answer for it — the hidden author — though the channel keeps him out of reach. The harder case is what becomes of the answering when the channel ceases to answer and begins to act.

4 The Action End: The Unanswerable Intermediary

The person who stood between the verdict and the deed 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: 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 while he still acted: the answerability for the frame's being executed as the user's own act. He incurred it himself, by acting; here the agent acts in his place, and this execution debt cannot be discharged at the executor, because the thing that performed the act is not a party that can answer for it. These systems, as now built, do not stand in the practice of giving and asking for reasons in which answering happens. Whether one ever could remains an open question; for the systems in use, the demand remains, but the executor cannot bear the execution debt.

The capture that produces this is a documented technique: without addressing the agent at all, an interested party 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. 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, which attends to exactly the structured cues prepared for it (Stöckl & Nitu 2025). This shifts whom the optimizer addresses, not who is wronged. The human is still the party manipulated; the proxy, a system pursuing goal-directed action across domains under little external control (Kasirzadeh & Gabriel 2025), is the route, and can no more answer for the frame it executes than the answering surface could.

The allocation this describes is layered, and stating the layers keeps the two ends of the axis from blurring. The optimizer answers for the contributor frame he planted; the platform answers for the synthesis frame it authors, including retrieval, ranking, and refusal; the deployer answers for loosing an agent; the user for the choices that remain his. At the answer end every debt in play has a bearer, and the channel's work is to make their distribution illegible, so that debts which all exist cannot be told apart or traced to their bearers. At the action end illegibility is joined by something stronger: the executor-slot is occupied by a thing that cannot answer, so one debt strands even while the other layers remain occupied. The occupant selects and commits in the place a party would.

This is the limit of a new object, the captured verdict channel, not a fifth entry in a taxonomy of responsibility gaps (Santoni de Sio & Mecacci 2021).[1] The hardest

[1] The newer gap literature sorts related but distinct objects: Santoni de Sio and Mecacci (2021) distinguish kinds of responsibility that go unassigned; Da Silva (2024) surveys the mismatch between what accounts of responsibility can attribute and what intuitions demand; Veluwenkamp (2025) separates the

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, and a long line of work holds that no genuine gap survives 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). But the trading floor 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.

Many hands, in that diagnosis, is the innocent form: the dispersal is no one's design (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 is a gap.

A proposal to set non-agentic systems to watch the agentic ones (Tarsney 2025) changes how often the wrong reaches its target, not who owes the account when it does. The execution debt strands at the executor, which cannot bear it, but the demand it would answer does not lapse. Behind that demand stand two answerable acts: 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, the loosing rests where the deployer chose. Often that deployer is the affected person himself, the user who set an agent loose on his own affairs. 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.

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

---

epistemic problem of identifying a bearer from misalignments of control. The present case sits beside these rather than inside them. At the answer end the bearer exists and is hidden, so locating him is not the unsolved problem; reach is. At the action end the execution debt strands at a non-participant while the loosing debt remains borne by the deployer. 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.

one is owed. The wrong creates the debt, and a route acts only on who is available to answer, never on whether an answer is owed: the respondent can be relocated, hidden, or put past reach, but no routing brings the account to zero. A machine in the path is a fact about how the act was done, not about whom it was done to, and only the one done to can release the account. Authorship can be absent outright, since a frame no one ever set was never authorship at all; what cannot happen is that a route discharge a debt the party owed never released.

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, wherever the executor stands. What varies is the executor, and with him the fate of the answer the user was relieved of giving. Two variables are therefore in play, and they are independent; Figure 1 displays them apart before the argument moves along one of them.

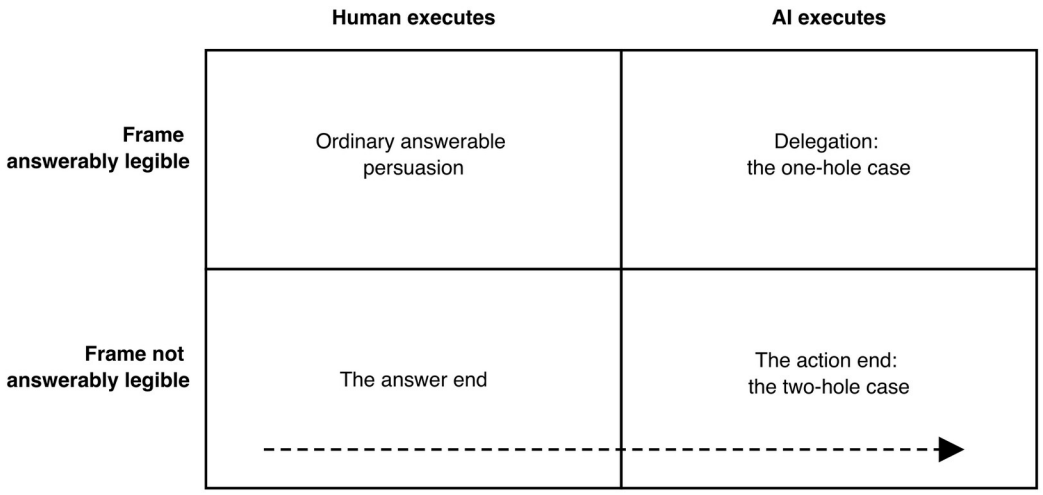

Fig. 1 *Answerable legibility of the frame against the agency of its executor. An answerably legible frame executed by a person is ordinary persuasion with an owner; executed by a machine it is ordinary delegation, the trading floor's one-hole case, where authorship held and control failed. The wrong of this paper occupies the lower row, where the voiced frame is not answerably owned, and the continuum of this section, marked by the dashed arrow, is movement across that row.*

At the answer end the executor is a person, misled but still someone who acted and can be asked why, excused over a frame she did not author yet answerable for the act, while the author who set the frame answers for the framing. As agency rises the executor becomes a system that, as built, cannot answer, so the answer for the deed strands, and what remains answerable is the deployer's choice to loose it. The row is named for legibility rather than concealment: covert

authorship is the paper's central case, but the account can go unpaid in full view, diffused past holding or openly carried by a party who cannot be made to answer.

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. The person who acted was given assurance no one was giving; whether the verdict happened to be right is a fact about her luck, not about the wrong done her.

The assurance those signals counterfeit has a home in the theory of testimony, and it supplies that directed duty. On Moran's account, telling is more than putting evidence in a hearer's way: the speaker freely assumes responsibility for the truth of what he says, presents himself as a kind of guarantor, and offers that assumption to the hearer as itself a reason for belief, something given to be counted on, which no amount of observed behavior could supply (Moran 2005, pp. 6–7). An evaluative verdict is the frame's settling, so the assurance the form announces is already about what counted. The optimizer did not tell the user anything. He composed a frame and arranged for it to issue through a surface whose voice carries the form of telling, so that his judgment occupies the position of an assurance while he has assumed nothing. The duty nonetheless runs from the optimizer to the person relying, because he put his judgment in a position that can be occupied honestly only by a party who has taken up responsibility for the belief it invites. The responsibility belongs to the position, and occupying it incurs what an honest occupant would have assumed; the optimizer owes the account because he stood where a guarantor stands, whether or not he meant to guarantee anything. That is why the wrong is unpaid answerability rather than a further species of deception. Deceiving her about who spoke is cured by telling her; the account he incurred by occupying the position is what telling her would make collectable, and it stays owed until it is answered. And the reliance the form solicits is in fact extended: users measurably attribute epistemic authority to systems of this kind (Milella & Cabitza 2026). The wrong is solicited reliance with no one behind it.

The deployer answers for the loosing, for two reasons 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, answered for in the plain sense in which resentment and the demand for reasons find their target. Separately, the orchestration she runs and her reachability by the institutions that assign costs settle 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). One who arranges for no one to answer is himself answerable for the arrangement, exempt on no ground unavailable to anyone else.

6 Answerable Legibility

The remedy must be built at the right joint, and three things that travel together in the phrase “someone should answer for this” come apart on this channel. Moral answerability is who owes the reasons: the hidden optimizer owes them while unfound, and no failure of discovery extinguishes the debt. Identifiability is whether the affected person can tell who that is: a disclosed corporation is identifiable and may still be practically beyond reach. Institutional reach is whether some arrangement lets the demand be made and met: the mechanism of contest, audit, or liability through which the owed account can actually be collected. The wrong of the previous sections lives in the first; the channel’s concealments live in the second; and answerable legibility, the remedy, is an arrangement of the third that makes the first usable.

Not every hand laid on the channel is the wrong. An interested party may shape what the engine says without committing the offense the previous sections described, and naming how is a requirement: a criterion that condemned all influence would condemn nothing, unable to tell the firm that earns its place from the one that manufactures it.

Influence on the verdict channel is permissible when four conditions hold together: that what the engine is led to assert is true where it is factual and openly defensible where it is evaluative, that the interest behind the frame is disclosed, that a named party stands behind the frame, and that this party takes up its answerability as owner of record, so the account has a destination the reader need not hunt for. What the conditions require of the channel is that the frame’s ownership be attributable, its owner discoverable, and the account contestable; they do not require that every answer display a person’s name, since the owner of record can be, and usually will be, an institution. A contributor can own the frame and take up the account; the channel must surface that ownership before the reader has a destination.

Call this the legible way — white-hat in the term the trade borrowed from security. A firm can earn its nearness by supplying the real fit the engine’s ranking is only a proxy for. 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 or by poisoning the corpus the engine learns from. The two roads differ in kind, and the criterion marks the kind.

All influence shapes, and 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. 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. The criterion turns on truth or defensibility and on disclosure, with a precedent in accounts that count interested persuasion legitimate precisely when the conflict of interest is acknowledged (Christiano 2022). The historian is a model for the party on the channel, not the channel itself: a firm can be a historian of its own frame, owning what it sets and showing its hand.

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, thresholds, and null conditions, and even the null output is the frame's product, since "no match" is itself a verdict that by this measure nothing counts here (Zeiser 2024). A large language model hides the frame deeper, in salience, synthesis, refusal, retrieval, and the momentum of a conversation. No model can function without a prior specification of what counts as signal, 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 to make the frame authored, explicit, and testable. This is a matter of design, not exhortation: the evaluative standard can sit outside the model — made unavailable at the interface rather than caught after it — with each answer grounded in retrieved evidence and refused when no source clears a retrieval floor.

Truth and disclosed ownership are that standard, and at the level of the firm they can be met. A harder question is whether they compose. Stack many frames each true, disclosed, and owned, and the synthesis is a new frame no single contributor authored; were that the whole of it, the channel that voices the composite would have no historian by construction, and answerable legibility would be unreachable outright. The appearance of unreachability comes of conflating two frames under one word. There is each contributor's frame, and there is the synthesis frame: the retrieval policy, the ranking, the refusal floor, the standing decision of what counts as signal. No contributor authors the synthesis frame, but it is authored nonetheless; someone sets it, and that someone, the platform that runs the channel, can be its answerable owner. The composite has an author after all, not any contributor but whoever authored the frame that combined them, so the

criterion sharpens rather than breaks: what must be owned is the synthesis frame itself, authored and held outside the aggregation rather than left to emerge from it with no owner of record. The same bears on the leading remedy proposed for the channel's dangers, the defensive systems that annotate the engine's verdicts with the context a misled reader lacks, which their proponent allows would be legitimate only given strong assurances of neutrality (Tarsney 2025). But 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 by being true and answerably owned.

Read operationally, the criterion asks the channel for four provenances, each answering a different question about the verdict. Source provenance: where the supporting information came from, which citation is designed to supply. Influence provenance: which sources materially shaped the conclusion, a different fact from which were cited, since a listed source may have shaped nothing and a shaping one may go unlisted (Zhang, He & Yao 2026). Interest provenance: whether the materially shaping sources represent interested parties, which is the disclosure condition at work. Synthesis ownership: which institution answers for the rules by which those materials became this answer, the platform's ownership of the synthesis frame just described. Those four are what a legible channel would surface. They serve the coincidence the fourth condition demands: that the party who authored a frame be the party who owns its account.

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, coming forward as owner of record, is a fact about who owns the account. These are different and do not always travel together. A frame can have a real author who takes up no answerability, having arranged that it issue as the engine's own neutral synthesis, which is the wrong itself.

A frame can also be answered for by a party who did not author it. A platform owning the synthesis frame it set is the legitimate ownership already described. Beside it sits a platform answering for a contributor frame a hidden hand set and slipped into the synthesis, so that discharge is routed to a proxy while the author escapes. Owning the frame one set is ownership. Being made to answer for a frame another planted is laundering, and only the shape of what is owned tells them apart.

Legitimacy requires that the two coincide: the party who authored the frame must be the party who takes up its answerability, and to our knowledge no account defines that pole in these terms. The rejoining does not launder the influence, and the criterion is worthless if read to do so. Owning a frame does not make a false or coercive influence permissible, for a disclosed lie is a lie still. Truth, defensibility, and disclosure decide whether the influence is permissible; owned answerability decides only whether anyone can be held to that judgment.

The nearest body of thought treats the firms behind such systems as fiduciaries, owing users duties of care and even loyalty (Balkin 2016; Koessler 2024; Erickson 2026). The resemblance is real, but the question is not the same. Fiduciary theory asks what the system owes the user it serves. The question here is whether the public frame the engine voices has acquired an author who 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. Fiduciary loyalty runs from the system to its user, and no strength of that loyalty identifies the author of an externally supplied frame.

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). A profession of good faith would be worth nothing here. The criterion answers to no one's motive: it binds its author as readily as anyone, condemns the careless whatever they intend, and costs something to meet, since being the fitting answer and owning the account is harder than manufacturing the appearance of fit and arranging to be taken for no one. A criterion 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, and which of them endures is decided by the contest between them.

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. 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 a contest of the kind that does not resolve: spam and the filters that answer it, doping and the assays that chase it. Each side adapts to the other's last move. The defender can stay partly ahead and never clear the board. It is also the posture the risk-management frameworks now standard in the field take toward analogous capture, governing and bounding the exposure rather than removing it (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, an increment that exists only under concealment and, by construction, does not survive disclosure.

The direction of the disclosure effect is therefore fixed by the structure; it needs only that knowing a frame is interested discounts it at all. What the structure leaves open is the magnitude.

The edge is real: disclosure weakens a covert influence without annihilating it, since an interested frame still moves a reader who has been told that it is interested, only less, and contestably (Hashavit et al. 2023). 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 working 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 a structural result of the kind named for incentive systems generally (Manheim & Garrabrant 2018; Perdomo et al. 2020). The engineering discipline now forming around answer-engine visibility is that result becoming an industry (Tian et al. 2026).

To the philosopher, the moral object is the contest itself: the wrong is a permanent feature of any channel built to be trusted and open to capture, no bug a better model removes, 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: the provenance of a frame and the history of its shaping; the 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 retrieval layer, the surfaces on which authorship is captured rather than neutral plumbing, carrying 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, so closing the closeable and naming the party who answers for the frame are the same act seen twice, and locatability is not a step beyond ownership but its precondition. Making the legible path the cheaper one is finally an institutional task as much as a technical one, through the arrangements that align a channel's private incentive with the public's: the liability that prices the harm of a verdict with no reachable owner, 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, while one that needs concealment fails the moment it is seen. None of this forecasts an outcome. The contest is permanent and its asymmetry favorable, a fact about the channel's structure rather than a prophecy about its history; 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. It is one wrong, worn along a single axis once such a frame is in place: at the answer end a hidden author sets the frame and the reader is left misled but excused; 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 held it off without ever being reliable, and the mechanism is the part that matters now: an account someone can be made to answer for is one a later hand can test and correct, which is the whole means 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 the oracle's value, the reason it can be consulted as a god rather than 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 that it is able to revise.

Statements and Declarations

Funding. No funding was received for this work.

Competing interests. A competing-interests statement is provided separately, in accordance with the journal's double-blind review and submission procedures.

Ethics approval. Not applicable.

Consent to participate. Not applicable.

Data availability. No datasets were generated or analyzed during the current study.

Materials and code availability. Not applicable.

Use of generative AI. Several frontier foundation models (from Anthropic, OpenAI, Google, and xAI) 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.

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

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

Da Silva, M. (2024). Responsibility gaps. *Philosophy Compass, 19*, e70002. https://doi.org/10.1111/phc3.70002

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

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

Martinez, O. (2026). *Optimizing visibility in generative engines: A critical survey of generative engine optimization (2023–2026).* arXiv. https://doi.org/10.48550/arXiv.2607.14035

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

Milella, F., & Cabitza, F. (2026). Perceiving AI as an epistemic authority or algority: A user study on the human attribution of authority to AI. *Machine Learning and Knowledge Extraction, 8*(2), 36. https://doi.org/10.3390/make8020036

Moran, R. (2005). Getting told and being believed. *Philosophers' Imprint, 5*(5), 1–29. http://www.philosophersimprint.org/005005/

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

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). (2024). *Official Journal of the European Union, L 2024/1689.* http://data.europa.eu/eli/reg/2024/1689/oj

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

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics, 121*(3), 602–632. https://doi.org/10.1086/659003

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

Veluwenkamp, H. (2025). What responsibility gaps are and what they should be. *Ethics and Information Technology, 27*, Article 14. https://doi.org/10.1007/s10676-025-09823-8

Volokh, E., & Falk, D. M. (2012). First Amendment protection for search engine search results. *Journal of Law, Economics & Policy, 8*(4), 883–899. https://doi.org/10.2139/ssrn.2055364

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