The Decision No One Authored

The Answerability Gap in Generative AI

Luke F. Walton

Independent Researcher · lukefwalton.com

ORCID: 0009-0005-9263-1954

Preprint. Not yet peer-reviewed. · v1.5 · July 2026

Companion to *The Captured Oracle: Authorship and Agency in the Ethics of Answer-Engine Optimization*

https://doi.org/10.5281/zenodo.20676327

Abstract. Debates over automating decisions in law, medicine, and public administration divide over what human judgment adds to the application of a rule. Both sides assume judgment's work is to supplement rule-application — to fill the gap between an abstract rule and the particular case. This paper identifies a prior gap the framing misses. Before any rule meets any case, someone must answerably settle what a system is for and what its outputs must meet: the evaluative frame the decision turns on. I call this *authorship*, the answerable setting of that frame and the bringing of each output under it. Authorship is distinct from, and prior to, the relations the leading governance concepts secure. Meaningful human control and human oversight can ensure that a system tracks a human's reasons, traces to her understanding, and remains within her capacity to override, yet leave untouched whether she authored the frame those relations operate within. The argument is analytic rather than empirical: a person can satisfy every control and oversight condition and still author none of what she accepts. Control governs her relation to the system's operation; authorship governs her relation to the frame through which operation becomes decision. I distinguish this answerability gap from the attributive gap the responsibility-gap literature addresses, show that it is invariant under the question of machine consciousness, locate the five junctures at which authorship is exercised or dropped, and argue that, when left unaddressed, the gap compounds rather than closes.

Keywords: answerability gap; authorship; meaningful human control; human oversight; moral responsibility; automated decision-making

1 Two Questions, Run Together

Suppose a firm adopts an AI system to assist its hiring. The system is trained on the firm's historical hiring records, configured to predict a target the firm labels "high performer," and deployed so that it returns a ranked shortlist from each pool of applicants. A recruiter reviews the ranking and decides whom to advance. The recruiter can override the system at any point; after the fact, the firm can say exactly which person signed off on which decision. By most ordinary lights a human being remained responsible for the hiring. And yet it is far from obvious that anyone, in the relevant sense, authored the decisions the system shaped. The recruiter inherited a definition of merit encoded in data she never examined, accepted a ranking as though it were evidence rather than a proposal, and treated a contested evaluative question, whom should we hire and on what grounds, as already answered by the time it reached her desk. She had control. It is much less clear that she exercised judgment.

The debate this collection takes up divides over what human judgment adds to the application of a rule: critics hold that sensitive decisions need a human to supply what no rule codifies, while enthusiasts answer that human judgment is just the noise a rule should discipline away. Both sides share an assumption — that the work judgment does is to *supplement rule-application*, to fill the gap between an abstract rule and the particular case. The hiring case exposes a gap of a different kind, one prior to the gap judgment is asked to fill. Before any rule is applied to any case, someone must have settled what the system is *for*, what would count as a good output, and that this question — whom to hire, on what conception of merit — is the one the machine will now answer. That settling is not rule-application and not the discretion that supplements it; it is the authorship of the evaluative frame within which rules and cases later meet. My claim is that this prior act is what goes missing when decisions are routed through fluent systems, and that the leading governance concepts — meaningful human control, human oversight — do not reach it. That is why a person can remain fully in control of a decision and still fail to author it.

A second conflation reinforces the first: public argument centers on the machine's status (whether it is conscious, sentient, the sort of thing that can be wronged) and supposes that settling it settles responsibility. It does not: the machine's status and the human's answerability differ in subject and, I will argue, in answer.

The literatures that have thought hardest about preserving responsibility (responsibility gaps, meaningful human control) can already connect a human to a system closely enough that an outcome is attributable to her. For the systems now in ordinary use, attributability is not the binding constraint: the recruiter in my example is attributable; the firm can name her. What has gone missing is not the locus of responsibility but its content: the evaluative work that makes holding her responsible accountability rather than a search for someone to blame. The norm that names that work is

what I will call authorship. The right question is not whether a human was in the loop but which actor set which juncture, with what authority, and whether that juncture was fixed upstream or reopened in the present case.

The argument is philosophical, but it arises from a practical problem familiar to anyone who designs and deploys these systems: how to build AI that increases what people can do without dissolving responsibility for what they do.

2 Answerability and Patiency Are Different Questions

By moral patiency I mean the property in virtue of which an entity can be wronged: whether there is something it is like to be it, in Nagel's (1974) phrase, such that what happens to it matters morally for its own sake. By answerability I mean the property of being a fit bearer of responsibility for an action or outcome, the standing to be asked to give reasons for it (Shoemaker, 2011). Whether a human or institution is answerable for a machine-mediated action is a question about us, not the machine.

The patiency question remains open. Current large language models are poor candidates: indicator-based assessment finds no strong evidence of consciousness (Butlin et al., 2023, 2026; cf. Seth, 2025), and their first-person reports, trained on human descriptions and reversible by reprompting, are weak evidence (Chalmers, 2023). The disciplined position (call it practical agnosticism) holds that no decisive barrier rules out future candidates and certainty either way is unearned.

The temptation is to draw the moral directly: current systems are probably not patients, so the only constraint is what we owe other humans. The inference is too quick: the questions lie on different axes, and the answer to one does not fix the other.

Machine ethics has long had the separation available: Floridi and Sanders (2004) argued that an artificial agent can be a genuine source of moral action without the mental states that would make it blameworthy, and treated agency and patiency as distinct roles. The lesson is the structural separation, not their constructive program: source of moral action, fit bearer of responsibility, and possible recipient of moral treatment come apart.

Full answerability requires more: to hold someone responsible, on Strawson's (1962) account, is to regard them as an apt target of the reactive attitudes (resentment, indignation, gratitude) extended to fellow participants in our practices of holding one another responsible. Responsibility in this sense is not having caused an outcome but standing in a relationship in which one can be addressed, can answer, and can be held to account. On current evidence the systems in use do not stand in that relation: to be angry at a language model, as opposed to angry about what was done with it, misdirects the attitude onto what cannot, as built, answer for itself.

Against that claim stands Dennett (1987): on the intentional-stance view, treating a system as a reasoner is not the discovery of an inner fact but a stance warranted by reasons-responsive competence, gradable and extensible in principle, so denying this system that standing can look assumed rather than argued. I grant the stance; the argument never required the machine to fall short. Predictive competence settles when we may treat a system as a participant in the space of reasons; it does not settle who authored the evaluative frame the decision turns on, because that frame was fixed about the system, not by it: the target it predicts and the history it learned from were settled before there was anything to take a stance toward. On either answer, the frame the model handed the recruiter was authored by no one: the missing act, not the missing participant.

Vallor and Vierkant (2024) converge from the responsibility-gap side, arguing that current systems lack the reciprocal standing our responsibility practices require; and the answerability at issue is not Tigard's (2021a) technological answerability, a system engineered to give answers, since saying why it produced an output is not being the party that must answer for its being acted on.

Two conclusions establish the decoupling. First, whether or not a system is a patient, none now in use is fit, on the evidence, to bear responsibility for a machine-mediated outcome in the place of the human who acted through it. A system that became a being we could wrong would not thereby become one we could blame in the recruiter's place, and one that became able to answer for its own acts would not thereby become able to answer for hers. Second, the human answerability the machine cannot absorb does not lapse because the machine is sophisticated, fluent, or surprising. It can be displaced and obscured by poor design; it cannot leave the human whose act it answers for, because that act is hers. Answerability is therefore invariant under the resolution of the patiency question. We do not need to know whether the machine has an inner life to know who must answer for what is done through it.

This is a stronger position than the one it replaces. "Current AI is probably not conscious, so humans remain in charge" is hostage to its premise; the decoupling grants the empirical point and denies the conclusion: answerability never rested on the machine's lacking a mind.

3 The Mirror and the Gap

If answerability cannot transfer to the machine, why is it so often lost? Three bodies of work triangulate the answer: two explain why abdication is tempting, the third why it is dangerous.

The temptation is illuminated by Vallor's (2024) account of contemporary AI as a mirror. Systems trained on the human archive do not stand outside our culture; they reflect our language, judgments, and institutional habits back to us, fluently enough to be mistaken for an independent source of insight. The diagnosis is right, and the fluency operates with particular force at the point of

judgment. A ranked list, a drafted paragraph, a recommendation: each arrives formatted as a conclusion, and the smoothness of the presentation is itself an argument for accepting it. The mirror does not merely tempt us to believe the machine understands; it tempts us to treat its outputs as if the evaluative work that would make them trustworthy were already done. Abdication therefore feels, from the inside, like reasonable reliance rather than surrender, a tendency well documented in operators' over-reliance on automated aids (Parasuraman & Riley, 1997).

Capability sharpens a turn the mirror account stops short of. Fluency tempts by making an output look finished; competence tempts by making deference correct. A reliably right system earns a trust the merely fluent one never could, the fit response to a track record. But a track record is also what dissolves the felt need to check, and most where the system has performed best. Vaughan's (1996) Challenger study located the failure in a normalization of deviance: a component performed outside its tested limits, returned intact, and each intact return read not as a near miss but as confirmation that the limit was safe to cross. Reliability had set the frame that no one authored. A competent system industrializes exactly this, one user and one uneventful Tuesday at a time: the better it works, the more reasonable it becomes to stop authoring, and the more complete the abdication on the case where the frame is wrong and nothing on the surface says so. The failure itself is not proprietary to generative systems; a regression can carry an unauthored frame as faithfully as a transformer. But generative fluency is what lets it scale and disappear: the more finished the output, the less visible the frame it inherits.

Those are the two faces of the temptation. The danger they court is the one Matthias (2004) made canonical: as learning systems become more adaptive, responsibility for what they do becomes hard to attribute by tracing back to a manufacturer's or operator's choices. A harm issues from data no one fully inspected, behavior no one fully predicts, reliance no one designed; the ordinary machinery of attribution finds no one who clearly intended or controlled it. Responsibility threatens to go missing between the human contributors, not to migrate into the machine.

Together they compose a single mechanism: the mirror supplies motive and cover, making abdication feel like good judgment; the gap supplies the cost, dispersing accountability across a chain of contributors until the judgment is no one's. The recruiter sits exactly at the junction: the firm can point to her, she to the system, the system to the data, the data to a history no present agent chose.

Some deflate the gap or deny it outright: Tigard (2021b) holds that our responsibility practices flex enough to encompass new technological agents, Königs (2022) that the circumstances of a gap are underspecified and its harms overstated, Demirtas (2025) that such gaps are neither new nor problematic, and Kasar (2025) that the difficulty is one of responsibility for unintentional action, traceable to human agents and modifying rather than negating their responsibility. Each is right

that nothing has migrated into the machine. But every such route reaches, at most, a locatable answerable party: a practice can always find someone to hold answerable without anyone having authored the decision. The question is therefore architectural rather than metaphysical (how should systems and institutions be arranged so answerability is preserved rather than dispersed?), and it is to the leading answer, and its limits, that I turn.

4 Why Control Is Necessary but Not Sufficient

The prior literature has approached this gap from several directions and stopped short of it in the same way each time. Accounts of the responsibility gap secure, or deny the loss of, *a human to whom the outcome can be traced*; decision-ownership requires that a decision-maker be positioned to *endorse* the values a system encodes; meaningful human control requires the system to *track* the relevant human reasons and *trace* to a human who understands it; human-oversight mandates require the capacities to interpret, resist, and override. Each secures a relation between a human and the *system's operation*. None secures the relation between a human and the *evaluative frame* through which that operation becomes a decision — the answerable exercise of judgment over what the system is for and what its outputs must meet. That relation is what I call authorship. The structure of the prior art's insufficiency is uniform: every one of these conditions can be fully satisfied while the frame the decision turns on was authored by no one. The existing accounts are not mistaken. They answer a different question, and the question they leave open is the one on which responsibility for a machine-mediated decision turns.

The most developed of these frameworks is meaningful human control. It originates in the autonomous-weapons debate, where Sparrow (2007) argued that no candidate party (programmer, commander, machine) could justly be held responsible for an atrocity. Santoni de Sio and van den Hoven (2018) gave it philosophical foundations: a system is under meaningful human control when it satisfies a tracking condition (responsiveness to the relevant moral reasons of the relevant humans and the relevant facts of the environment) and a tracing condition (traceability to the appropriate moral understanding of at least one human agent in its design or use).

This is a genuine advance over the slogan of keeping a human "in the loop," whose mere presence is a formality that licenses the institution to claim a human was responsible; Green (2022) presses this into a general indictment of oversight mandates as false assurance. The failure isolated here is the complementary one: the human can intervene, is fairly attributable, and the judgment over the frame is missing.

Developed against systems that act, the framework has since been operationalized for decision-support systems of exactly the hiring kind (Cavalcante Siebert et al., 2023); my claim is not that the extension fails but that, even succeeding, it secures control over an evaluative frame rather

than authorship of it. The recruiter can override any recommendation, the system tracks the firm's stated reasons, and its behavior traces to humans who understand it: both conditions are satisfied, and the failure described in §1 has nonetheless occurred.

Neither condition captures what has gone wrong, because both take the frame as given. Tracking is silent on whether the tracked reasons were the product of judgment or were inherited, unexamined, from the vendor's choice of target variable: someone decided that "high performer" was the thing to predict, converting a contested normative question into a settled prediction problem, a decision nowhere registered as requiring justification. Tracing to a human who understands the system is not judgment about what its ranking should mean.

This may understate the framework: tracking does not require responsiveness to just any reasons the firm holds; it requires the relevant moral reasons. If the target variable encodes a conception of merit corrupted by inherited bias, tracking is not satisfied after all, and meaningful human control already condemns the hiring system.

I grant the charitable reading. Tracking and tracing, however construed, are relational: they hold or fail between a system's behavior and a frame supplied from elsewhere, and say nothing about the act of supplying it. Authorship is that act, the answerable exercise of judgment that determines which reasons are the relevant ones, and to fault the system for tracking a defective conception of merit already presupposes a better conception someone was answerable for authoring. The point holds even if the relevant reasons are fully objective, for objective reasons do not operationalize themselves: between whatever merit really is and the target a system predicts stand choices of proxy, threshold, and scope that the true reasons underdetermine, a fortiori if reasons can underdetermine even which action they favor (Kozlovski, 2025). The realist needs an author no less than the constructivist — not to make the values true, but to make some determinate conception of them operative and answer for the making. The determination still owed is an authorial act, not a refinement of control.

The two notions come apart in both directions, which is what shows them to be distinct norms rather than one norm at two strengths. A frame can be defensible and yet unauthored. Consider a triage system that orders incoming support tickets by predicting which will escalate, on a definition of escalation no one disputes. Tracking and tracing are uncontroversially satisfied even on the most demanding reading: the system answers to a reason everyone endorses, and its behavior traces to operators who understand exactly what it does. And yet the operator may have authored nothing. She accepts the ordering without having decided what standard a prediction must meet before she acts on it, and without ever making the ordering her own judgment rather than the system's default. The frame is correct and the authorship is hollow. Conversely, a frame can be authored and yet wrong, since authorship tracks the answerable exercise of judgment, not the correctness of its result. The framework's own authors grant as much: in their example of a commander who

knowingly deploys an autonomous weapon that cannot comply with the laws of armed conflict, Santoni de Sio and van den Hoven (2018) note that "not only the tracing, but also the tracking condition is satisfied," even as the attack is unlawful and the commander culpable. Both conditions are met and the ends are still wrong. The absorption objection succeeds only by quietly identifying authorship with getting the values right; once that identification is refused, the distinction is secure, because correctness is neither necessary nor sufficient for authorship. Authorship is therefore not thicker control but a prior condition that control presupposes and does not supply.

Control, when it succeeds, secures a locus of responsibility but not its content: whether the connected human performed the evaluative work in virtue of which the attribution is just. Two failures hide under the single phrase "responsibility gap." The first is attributive, the gap Matthias and Sparrow identified: no human is connected closely enough to bear responsibility at all. The second is an answerability gap: someone, often many people, is attributable for the outcome, but no one exercised the answerable judgment over its evaluative frame that responsibility is supposed to track. The distinction is Shoemaker's (2011), between attributability and answerability, here sorting gaps in a distributed act rather than kinds of one agent's responsibility.

The nearest decision-support account, Zeiser's (2024), isolates a problem of decision-ownership — what he frames as preserving human authorship in decisions — from a closely parallel hiring case and Shoemaker vocabulary: a decision is owned insofar as it reflects the decision-maker's value-judgments, which she should be able to explain and answer for. What separates the present account is his leading repair: being positioned to endorse the value-judgments the system presupposes. Endorsement-capacity is a standing relation, not an occurrent act: it asks whether the agent would endorse a frame supplied from elsewhere, not whether she set it — endorsement on request being, §5 will argue, the mark of a frame inherited rather than authored. The gap Zeiser opens is real; the capacity to endorse does not close it.

Despite the name, the answerability gap is not a shortage of attributable parties: in the hiring case they abound (firm, procurer, vendor, recruiter) and the wrong is not that responsibility lacks somewhere to land but that none of them performed the answerable judgment it presupposes. Kiener (2025) argues, against the gap tradition, that AI-mediated harms typically leave too many attributable parties, not too few. The abundance is real but orthogonal: it is a fact about attribution, and a crowd of attributable parties can each have taken the frame ready-made from the next — deployer from vendor, user from deployer. Abundance is the complement of the answerability gap, not a rival. A genuine gap exists only when no juncture was occurrently authored: target from convention, convention from data, data from a history no present party chose and owns; where some juncture was authored but blame lands elsewhere, the failure is a misallocation of attributable authorship, nearer many hands than a gap. The recruiter's is the first kind, even where the vendor deliberated over "high performer" before settling on it: a frame is more specific than a product's

general target, and what went unauthored is its fitness as the operative standard for this firm's hire of this role, a choice no one confronted and owned.

When the locus is secured but empty, holding the connected human responsible is closer to scapegoating than accountability: Elish's (2019) moral crumple zone names the operator faulted despite too little control; here the operator has control in abundance and what is missing is judgment, because the fluency that makes intervention unnecessary also makes abdication invisible. And the zone is not assigned at random: it is engineered into the cheapest, most replaceable region, so when a frame authored by no one comes due, the account runs downhill onto the approver who did not procure the system, configure its target, or choose the vendor — the party the same institutional logic can afford to lose.

The problem of many hands (Thompson, 1980) is usually told as diffusion; but the diffusion is not symmetric and does not net to no-one-pays. It strips answerability from the parties with the standing to shed it and deposits the remainder on the party with the least: agency laundering (Rubel et al., 2019), then the laundered account coming to rest. "Human in the loop" names a safeguard, but a loop is also the shape a blame circuit takes, a station wired in so the circuit can close, when an outcome must be answered for, on whoever occupies it — an occupant where an author was owed; and the position is a position, not a person, refilled when its occupant is consumed. The account does not vanish into the crowd of hands. It moves: never discharged by the parties who incurred it, never forgiven, only relocated, until it rests on whoever is least able to move it on. Whether a wrong always leaves such an account, invariant under mediation, is more than one case can establish; I take it up separately. What the case forces is enough: the answerability gap is not dangerous because blame goes nowhere. It is dangerous because blame goes somewhere predictable: downhill.

The rotation has a second effect, on the frame rather than on the person, and it is the more lasting of the two. When the recruiter leaves and another fills the chair, the unauthored frame is handed on intact, and the next occupant inherits it exactly as she did, as a given, not a choice. Across enough cycles the frame is not merely unauthored once but continuously: it drifts, as the system is retrained on its own accumulating outputs and as each occupant ratifies what the last left in place, and at no point does any party confront the drift as a decision. The result, after some years, is a conception of merit the firm now runs on that no human ever chose and none can account for — not because the record was lost but because the authoring was never done. Whether the drift is large or small is an empirical question and beside the point; what is structural is that there is, by construction, no one positioned to say why the frame is what it has become, or to be asked to correct it. This is why the answerability gap is not a momentary lapse but a ratchet. A frame no one authored cannot be answerably revised, because revision too is an authorial act, and the same conditions that prevented the first prevent every correction after. The abdication compounds, and

what compounds is precisely the absence of anyone who could arrest it.

The failure is institutional first: ends and standards were settled upstream, in the choice of vendor, configuration, and target, and that is where the missing authorship was owed. Nor is the gap among the four Santoni de Sio and Mecacci (2021) distinguish: theirs sort kinds of responsibility a party may fail to bear; this one cuts within secured attribution, nearest their active responsibility without being a species of it. The recruiter is positioned to author going forward, and the gap opens exactly there, between being positioned to author and having authored. Arendt (1963) named the local failure thoughtlessness: action continued without the reflective interruption in which one asks what one is actually doing. What transfers is the concept: a person can be present, capable, and compliant while having suspended the judgment that would make the action hers. The next section develops authorship as that interruption made a standing requirement rather than a private virtue.

5 Authorship

I define authorship as the exercise of answerable judgment at five junctures of a machine-mediated action — two at which its evaluative frame is set, and three at which each output is answerably brought under that frame: the ends the system serves, the standards by which its outputs are evaluated, the conditions under which those outputs are verified, the moment at which an output is accepted into action, and the final form for which someone stands answerable. The first two set the frame; the last three are exercised anew with each decision, and they are not further frames but the acts by which each output is answerably brought under the frame already set: measured against the standard, admitted under it into action, owned as it goes out. Nor are they control under another name. Control concerns the capacity to intervene in the system's operation; these three junctures sit where operation becomes action, and what they demand is not the capacity to intervene but answerability for the admitting — a judgment, exercised on this case, that the standing frame governs it and is met. Authorship is not solitary creation: the author need not perform any task the system performs, provided its use remains answerable at the five points; for deployed systems that author is human or institutional, and the requirement is on the authorship, not the author; whether a future system could itself answerably author is the question §2 left open.

The five are not a checklist of governance desiderata but the junctures at which a single relation, answerability for the action, is discharged or dropped, and each is individually necessary: a decision taken toward unexamined ends is not redeemed by careful verification, and a well-framed decision accepted by default is unauthored at the moment it becomes an act. Authorship admits of degree along these dimensions; the five are defeasible constitutive conditions for responsible machine-mediated action.

The term is used here in a restricted, responsibility-centered sense, for answerable judgment over

a decision's evaluative frame: not in the aesthetic or proprietary sense, nor the credit-and-blame sense in which Nyholm (2024) asks how praise or blame for generative outputs attaches to the user.

The everyday sense runs the other way. A model that drafts a paragraph has, in ordinary speech, authored it; but making is execution, which transfers to a system without remainder; what does not transfer is answerability for the frame the artifact serves. A system can supply every word and answerably author none of it, just as a person can answerably author a decision whose every word a system supplied. When I say a frame was authored by no one, I mean that no party answerably owned it, not that no party shaped it.

Nor do I claim to have discovered the junctures. Several appear, in operational dress, in the EU AI Act's Article 14, which requires that those overseeing a high-risk system be able to understand its limits, resist automation bias, interpret its output, and decide not to use it (Regulation (EU) 2024/1689). But Article 14 secures capacities for oversight; authorship is the relation in which exercising them preserves responsibility. The contribution is the organization, and the relation it serves, not the bare list.

Authorship of ends. Every deployed system is for something, and what it is for is a value choice the system cannot make. Whether a hiring tool should predict tenure, productivity, or a richer conception of what the firm should want in its people is not a technical question. The choice hides inside an objective or target variable, making a normative decision look like a modeling one; translating an aim into a target is a documented site of discretionary, normatively loaded choice (Passi & Barocas, 2019). To author the ends is to recognize the value choice and own it rather than accept the objective that arrives pre-installed.

Authorship of standards. Distinct from the ends a system serves is what counts as a good output. A model can apply a standard, but it cannot be the source of the standard by which its own output is judged, on pain of a circularity in which the system both produces the work and certifies it. The criteria must be specified outside the model, by parties answerable for them, so the output is measured against something it did not generate; where standards are tacit, or read off the model's own confidence, authorship of standards has lapsed even with every other safeguard in place.

Authorship of verification. A fluent output looks finished, the mirror's characteristic danger. Verification reconnects an output to the reality it purports to be about: the claim checked against the source, the summary against the document. Competence sharpens the danger, since each past success is genuine evidence about the next; verification is the juncture a competent system erodes first, and fastest where it has performed best. What must be checked, against what, and before which consequences may follow is fixed with the standards; what verification authors is the performed check — that the checking actually occurred, on this case, as this party's act rather than an assumption.

Authorship of acceptance. In any machine-mediated action there is a moment at which a proposal becomes a deed: the draft becomes the sent email, the ranking the rejection letter. This is the hinge of authorship, where the human either exercises judgment or merely ratifies, and the deepest design failures collapse it by making acceptance the default or the path of least resistance. The limiting case is familiar from coding agents that offer auto-accept modes: settings that run commands and write files without a confirmation step, colloquially called "YOLO mode." It is the acceptance juncture removed by design.

Authorship of final form. Authorship includes standing behind the artifact as it goes out: answerability not only for releasing it but for what it does, including consequences unintended but foreseeable. It begins before any harm, in the willingness to be the one who answers if it is questioned.

It helps to see the five satisfied rather than only breached. Consider a physician using a system to draft a clinical note. The end is one the physician has authored: an accurate record in the service of care, not throughput. The standards are external to the model, held by the profession: clinical accuracy, completeness, the norms of the record. Verification is built into the act: she reads the draft against what happened in the room and corrects it. Acceptance is explicit: nothing enters the record until she signs. And the final form goes out under her name, hers to answer for. The system does a great deal, yet authorship is intact: its output remained a proposal, measured against standards it did not set, by the person who answers for the result. The contrast with the hiring case is not that one uses AI and the other does not; both use it heavily. The difference is whether the evaluative frame was authored or inherited.

Two features protect against predictable misreadings. First, authorship includes control but exceeds it: much of what a system does (retrieval, formatting, computation, synthesis) can be handed over entirely without loss of authorship, provided the five junctures remain answerably human; authorship is not a brake on automation but a constraint on where it may run without remainder. Second, authorship is an exercised relation, not a location: "human in the loop" is the wrong unit (Crootof et al., 2023), since a human can occupy the loop while authoring nothing, and answerability can hold from far upstream if she authored the ends and standards in force.

What distinguishes a frame a human has authored from one merely inherited and waved through? If her conception of merit is itself absorbed from her culture, the "authorial" hiring manager looks no less a conduit than the model, and the demand threatens to regress without end.

The regress dissolves once one declines a premise it smuggles in. A model's relation to its corpus and a person's relation to her culture are not the same relation: the model's outputs can be steered into a different answer, but no one is addressed by the steering; a person can be asked why this conception of merit and not another, and can change it, because the question lands on an agent who must answer. That is the asymmetry — not malleability, which both have, but answerability,

which only one of them has. The objection tests authorship against origin, but authorship was never origination: an absorbed frame can still be authored.

And holding a frame answerably is something a person must do, not something she would do if asked. Someone must have confronted the value choice the frame encodes, recognized it as a choice, and made it; a disposition to defend a frame one never set is the mark of not having set it. Occurrence is a fact a record can show, not an introspective episode: a default examined, a target contested, a standard set where none was supplied — the closing locatable as someone's doing. The exercise is occurrent, but it is not therefore constant: the frame is authored once, by whoever sets it — possibly far upstream — and the per-output acts anew with each decision. This is why speed does not defeat the physician: her profession occurrently set her standards, and she occurrently verifies and accepts each note, however fast she signs.

Nor is the profession's authorship merely the recruiter's inheritance one level up. The difference is an act with a name: adoption. The profession authored its standards as standards, and the physician adopted them by an answerable act, entering a practice that holds her to them; what remains open per encounter — the fit of the standard to the case before her — is exactly what her verification and acceptance answer for. Authorship of ends and standards can thus be discharged far upstream, by answerably adopting what an answerable party set; adoption is occurrent, a confrontation that closes in someone's name. What persists thereafter is not a disposition doing authorial work but the standing effect of a completed act, as a promise binds after the promising; what mere endorsement-capacity lacks is not persistence but occurrence. What it cannot be discharged by is procurement. The firm bought a tool containing a definition of merit, and no one made that definition the firm's own: no adoption upstream, only ratification at the desk, so the fitness of "high performer" as the operative standard for this hire was confronted nowhere. And adoption is what the regress objection was asking for — the act by which a frame absorbed from elsewhere becomes answerably one's own.

The occurrent requirement is also why the recruiter fails despite passing every dispositional test. Asked, she would defend "high performer," hear an objection, perhaps revise; but no one occurrently authored the ends her system serves, and standing willingness to justify what one merely received is not authoring it. Reflective deference fails the same way: "I have considered it and I trust the system" makes the trust answerable while leaving the frame exactly as mute as she found it. A long tradition ties asserting to a standing readiness to defend what one has put forward (Brandom, 1983), but that commitment runs to the product, the claim asserted, whereas authorship concerns who set the frame: one can be fully answerable for defending a recommendation while having authored none of the choices that made it.

Authorship cannot pass to the model for a reason that survives every improvement in its training. What a system has are sources only in the sense of provenance — where its material came from —

not in the sense of backing, a party who stands answerably behind a claim and is accountable to a reality it might get wrong. The distinction matters because the machine has no reality it registers as its own to answer to: when it is corrected, the correction lands on the next output, not on anyone who must say why the last was wrong. A person tracking such a reality registers constantly the gap between what her sources settle and what they do not — the Socratic mark of knowing the limits of one's knowledge — and that posture is precisely what authorship requires and what a generative system, as built, cannot take up. A perfectly curated, perfectly current corpus would change none of this. Authorship needs an agent who can answer for the frame, and curating data supplies no such agent.

Authorship, it will be objected, is the romantic fiction of the solitary maker; real institutions distribute responsibility across many hands. The objection mistakes distribution for dissolution. A film has hundreds of contributors and a structure of authorship nonetheless; authorship can be distributed across the five junctures, provided each act was someone's answerable judgment and the chain reconstructable; the bearer may even be an institution (List, 2021). What converts distribution into dissolution is the absence, at one or more junctures, of any hand at all. Distributed authorship is how complex action is properly governed. Dissolved authorship is the answerability gap under another name.

A final objection concerns demandingness. If authorship requires that someone occurrently confront the value choice a frame encodes, much institutional life before AI fails the standard too; so either the answerability gap is everywhere and the machine is incidental, or the standard quietly relaxes whenever it would convict ordinary practice. I refuse both horns. The standard never relaxes: it was always failable, and this paper's own diagnostic cases predate the technology; normalization of deviance and bureaucratic thoughtlessness are old failures of exactly this kind, convicted by the same criterion. Nor is the gap everywhere: adoption shows the standard ordinarily met, and the physician satisfies all five junctures at working speed because her profession confronted the value choices once, answerably, upstream. And the machine is not incidental: what generative fluency changes is not the standard but the economics of failing it, the scale at which inherited frames are installed and the invisibility of their acceptance. The gap is not everywhere, and it is not new. What is new is how cheaply it is produced and how little it shows.

6 Three Implications for Design

If the binding problem is abdicated judgment rather than absent control, design should be evaluated by whether it preserves authorship, not whether it inserts a human somewhere in a process. Three implications follow from the five junctures of §5.

The first is to separate generation from acceptance. A system may generate a candidate output,

but the act by which it becomes action must remain distinct and identifiable, performed by an answerable party under standards fixed in advance. Resist designs in which acceptance is the default: the draft that sends itself, the ranking acted on unless someone intervenes. Proposals should stay legible as proposals until someone accepts them, and acceptance should be an act, not the absence of an objection: authorship of acceptance built into the architecture rather than left to the user's discipline.

The second is to keep the standards outside the model. The criteria by which an output is judged must not come from the system whose output they govern, nor silently from the model's own confidence or training distribution. They must be specified, and owned (adopted as the institution's own rather than inherited as a vendor's default), by parties answerable for them, held where the model can be measured rather than left to certify itself.

The third is to make responsibility traceable across layers. Because authorship is distributed, the defense against its dissolution is the ability to reconstruct the chain: who set the goal, chose the model, framed the interface, verified the outputs, accepted the consequences. A system for which this reconstruction is impossible has not eliminated responsibility; it has hidden it, which is worse. The design literature is converging on such commitments (Zhu et al., 2026); the contribution here is the answerability they presuppose, not the layering.

A limit must be owned; it is the thesis applied to its own remedy. These commitments can be implemented at the architecture level, disallowed operations made structurally inexpressible rather than prevented by convention. What architecture cannot do is manufacture the judgment: a target variable shown in bright letters can still be waved through, an inserted check clicked past as reflexively as it was added. Architecture can refuse to let the exercise of judgment be invisible. It cannot perform it for the user.

A caution cuts across the three: a system framed as a colleague or an oracle invites the deference that hollows authorship; framed as an instrument, it keeps the user in the authorial position. Anthropomorphic role assignment is not cosmetic: it partly determines whether the human will author what the system does or merely accept it.

7 Objections

Is this not Vallor's mirror thesis restated? No. The mirror is a diagnosis: it explains why deference to AI outputs is tempting. I accept it and build downstream of it, in three claims the mirror does not make: that human answerability is independent of the machine's moral status, that control is insufficient for the failure mode generative systems produce, and that the repair is a specific, design-relevant norm of authorship. Diagnosis and remedy are different achievements; it is the remedy I

am defending.

Is the gap not simply the familiar one between rule and case, the discretion particularism defends? No: discretion operates within a frame already set, choosing how a rule meets a case. Authorship is the prior act of setting what the system is for and what its outputs must meet. A casuist judging the particular still presupposes that someone authored the ends the judgment serves; the gap named here is upstream of the rule-following the debate concerns.

Does the decoupling fail if AI becomes conscious, or a genuine participant in the space of reasons? It is built to concede both. Patiency would change what we owe the system; participation would let it answer for its own acts; neither makes it the bearer of the human's answerability for her act: earning the standing to answer for one's own conduct is not exercising judgment over the terms of someone else's decision. And a future system built so that no one answers for what it does is a fact about how it was built — an indictment of building it that way, not a discovery that answerability has migrated into silicon.

Does privileging human judgment ignore that it is itself biased? It strengthens the account: the hiring system's bias was inherited from human history encoded in the data, and the claim is not that human judgment is reliable but that responsibility requires an answerable party who exercised it. The right response is to author more carefully, not to abdicate to a system that launders the same bias while removing the one who could be held to account.

Does alignment dissolve the gap? To align a system is to make its ends good, not to make them anyone's. Even intent alignment, which makes the ends the principal's, secures a tracking relation; and tracking, §4 argued, holds between behavior and a frame supplied from elsewhere. Alignment is necessary and insufficient in just the way control is: an aligned model leaves the gap exactly where a misaligned one does.

Is this, in the end, a counsel against using AI? The opposite. The danger is abdication, not automation; a system whose users author the ends, standards, verification, acceptance, and final form of what it produces is better, not diminished.

8 Conclusion

Whether these systems are, or might become, minds we could wrong is a real question I have not settled; I have argued it is the wrong question to center, because its answer, whatever it proves, leaves human answerability untouched. Patiency is uncertain and may change; answerability is immediate and does not. We can usually name the human who signed off. What we cannot assume is that she authored what she signed.

The deepest temptation the mirror presents is not that we will mistake the machine for a person,

but that we will let it do our judging and call the result a decision. Fluency opens that door and competence walks us through it, for the reason §3 gave: reliability earns a trust that makes not-checking reasonable, and the reasonable abdication is the hardest kind to interrupt, because nothing about it feels like surrender. The discipline this requires is the reflective interruption described earlier: the moment in which a person asks what is actually being done, and whether she can answer for it. To build and use these systems well is to keep that interruption alive, to ensure that somewhere in the chain between the model's output and the act it becomes, a human being remained answerable for the judgment the machine was used to make. Neglecting the answerability gap extends the separation of liability from authorship, until someone answers for everything and authored none of it. The likelihood of such a future increases as the machines improve. Every increment of capability is an increment in the rational case for deference, and so in the ease of the abdication: the better the system, the less anything feels wrong as the judgment quietly stops being anyone's.

Competing interests

The author is the founder of Surmado, Inc., which builds AI-orchestration systems for small businesses. This work was conducted in the author's personal capacity; the views expressed are the author's own and do not represent those of any employer.

Funding

No funding was received for this work.

Use of generative AI

Four frontier foundation models supported literature search, section-level drafting, argument pressure-testing, revision, and formatting: ChatGPT 5.5 (OpenAI), Claude Opus 4.8 and Claude Fable 5 (Anthropic), and Gemini 3.5 (Google). The author originated the thesis and its central distinctions, set the standards for inclusion, drafted sections of the text, directed and revised all model-drafted material, and verified the claims, quotations, and citations against primary sources rather than against model agreement. The author is answerable for the final form.

References

Arendt, H. (1963). *Eichmann in Jerusalem: A report on the banality of evil.* Viking Press.

Brandom, R. (1983). Asserting. *Noûs, 17*(4), 637–650. https://doi.org/10.2307/2215086

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences, 30*(6), 488–501. https://doi.org/10.1016/j.tics.2025.10.011

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness.* arXiv. https://doi.org/10.48550/arXiv.2308.08708

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics, 3*(1), 241–255. https://doi.org/10.1007/s43681-022-00167-3

Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review.* https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Crootof, R., Kaminski, M. E., & Price, W. N., II. (2023). Humans in the loop. *Vanderbilt Law Review, 76*(2), 429–510. https://scholarship.law.vanderbilt.edu/vlr/vol76/iss2/2

Demirtas, H. (2025). AI responsibility gap: not new, inevitable, unproblematic. *Ethics and Information Technology, 27*, 7. https://doi.org/10.1007/s10676-024-09814-1

Dennett, D. C. (1987). *The intentional stance.* MIT Press.

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

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*, Article 105681. https://doi.org/10.1016/j.clsr.2022.105681

Kasar, P. (2025). There is a problem, but not a responsibility gap. *Ethics and Information Technology, 27*, 47. https://doi.org/10.1007/s10676-025-09851-4

Kiener, M. (2025). AI and responsibility: No gap, but abundance. *Journal of Applied Philosophy, 42*(1), 357–374. https://doi.org/10.1111/japp.12765

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

Kozlovski, A. (2025). Reasons underdetermination in meaningful human control. *Ethics and Information Technology, 27*(4), Article 59. https://doi.org/10.1007/s10676-025-09858-x

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology, 34*(4), 1213–1242. https://doi.org/10.1007/s13347-021-00454-7

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450. https://doi.org/10.2307/2183914

Nyholm, S. (2024). Generative AI's gappiness: Meaningfulness, authorship, and the credit-blame asymmetry. In A. Strasser (Ed.), *Anna's AI anthology: How to live with smart machines?* (pp. 167–194). Xenomoi Verlag.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)* (pp. 39–48). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287567

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, 12 July 2024. https://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

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

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, Article 15. https://doi.org/10.3389/frobt.2018.00015

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences.* Advance online publication. https://doi.org/10.1017/S0140525X25000032

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

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48*, 187–211. https://www.thebritishacademy.ac.uk/publishing/proceedings-british-academy/proceedings-volumes-1-111/48/strawson/

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

Tigard, D. W. (2021a). Technological answerability and the severance problem: Staying connected by demanding answers. *Science and Engineering Ethics, 27*, Article 59. https://doi.org/10.1007/s11948-021-00334-5

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

Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking.* Oxford University Press.

Vallor, S., & Vierkant, T. (2024). Find the gap: AI, responsible agency and vulnerability. *Minds and Machines, 34*(3), Article 20. https://doi.org/10.1007/s11023-024-09674-0

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA.* University of Chicago 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

Zhu, L., Lu, Q., Ding, M., Lee, S. U., & Wang, C. (2026). Designing meaningful human oversight in AI. *AI and Ethics, 6*, Article 286. https://doi.org/10.1007/s43681-026-01147-7