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 Not yet peer-reviewed. · v1.4 · June 2026

Abstract. Can a person stay fully in control of an AI-shaped decision and still fail to author it? They may hold the power to override the system, be expected and encouraged to do so, and remain answerable for the outcome after the fact, yet never exercise judgment over the evaluative frame the decision turns on, inheriting that frame rather than authoring it. This article argues that the gap this opens, rather than the question of whether AI is conscious, is where responsibility for machine-mediated action is won or lost. The system's fluency is what widens it: confident, plausible, finished-looking outputs invite a person to accept the machine's framing wholesale, and that temptation operates whether or not anything is actually home. And the temptation deepens rather than recedes as the systems improve, since a reliable track record makes deference rational and erodes the felt need to check at the very juncture where checking still matters. The claim concerns responsibility, not mind: the gap is a missing act, not a missing participant, and it stays open however the machine's status is settled.

The dominant design-facing norm, meaningful human control, can close what I call the attributive gap, securing some human to whom an outcome can be traced. It leaves open an answerability gap: the traceable party, even amid a crowd of attributable parties, never exercised the judgment over the decision's evaluative frame that separates holding someone responsible from scapegoating them. Generative and decision-support systems are especially prone to produce this second gap, because their fluency lets a person keep full capacity to intervene while authoring none of what they accept.

The norm that closes it is authorship: the answerable exercise of judgment over a decision's evaluative frame, discharged at the five junctures where that judgment is most easily skipped. These are the ends a system serves, the standards its outputs must meet, their verification, their acceptance into action, and the final form for which someone stands answerable. The contribution is this positive account, and its reframing of answerability as a relation that must be exercised rather than a position one occupies. Authorship holds independent of the machine's moral status: it is owed whether or not the system is conscious, and whether or not it is right. It is likewise independent of

who can supply it: present systems require a human or institutional author, a sufficiently capable future system might not, and the requirement does not vary with the answer. More fundamentally, control governs the human's relation to the system's operation; authorship governs the human's relation to the evaluative frame through which that operation becomes action.

Drawing on Acemoglu and Restrepo's task framework and Acemoglu and Johnson's critique of so-so automation, I argue that the deeper danger is a so-so automation of judgment, evaluative work displaced without new tasks in which human agency is strengthened. I close with three implications and a caution for design.

Keywords: artificial intelligence; moral patiency; moral responsibility; responsibility gap; answerability gap; meaningful human control; answerable authorship; automation; philosophy of technology; AI ethics; Vallor; Acemoglu; AI alignment

Disclosures

Competing interests. The author is the founder of Surmado, Inc., which builds AI-orchestration systems for small businesses. The paper's diagnosis — including its treatment of answer-engine optimization as an emerging economics of the generative channel (§5) — applies to that commercial work as much as to any other actor's. No funding was received for this work, and the analysis was neither commissioned nor reviewed by any commercial party. 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.

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

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; the firm can say, after the fact, 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.

This article is about the gap that case exposes, and about why so much discussion of artificial intelligence misses it. The trouble begins with a conflation of two questions that ought to be kept apart. The first is a question about the status of the machine: is an AI system the sort of thing that can be wronged, that has interests or experiences of its own, such that we owe it moral consideration? The second is a question about the answerability of human beings: when something is done through an AI system, who is responsible for the ends it served, the standards it met or failed, and the consequences it produced? These questions differ in subject and, I will argue, in answer. The first is the question that dominates public argument, where it appears as the debate over machine consciousness, sentience, and moral patienthood. The second is the question that actually governs whether the hiring system, or any deployed system, is used well or badly. The two get fused because the machine's fluency invites it: a system that talks as though it understands, judges, and decides makes it natural to ask whether it really does, and natural to suppose that if it does not, the matter of responsibility is thereby settled in our favor. Neither supposition survives scrutiny.

The literatures that have thought hardest about preserving responsibility, the work on responsibility gaps and on meaningful human control, can already connect a human to a system closely enough that an outcome is attributable to her. For the generative and decision-support 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 substantive evaluative work that makes holding her responsible an exercise of accountability rather than an exercise in finding someone to blame. The norm that names that work is what I will call authorship. Authorship is prior to control rather than a stronger form of it, and that difference is what the hiring case turns on. The right question, then, is not whether a human was in the loop but which actor set which juncture, with what authority to frame the space of reasons, and whether that juncture was fixed upstream or reopened in the present case.

Although the argument that follows is philosophical, 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 the 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, the having of welfare or experience such that what happens to it matters morally for its own sake. The question whether an AI system is a moral patient is the question whether there is something it is like to be it, in Nagel's (1974) phrase, and whether that grounds claims on our treatment of it. 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). The question whether a human being or institution is answerable for a machine-mediated action is a question about us, not about the machine.

The patiency question is the one the public argument is really conducting, and it remains open. Current large language models are poor candidates for patienthood: base models neither maintain themselves as organisms nor inhabit a world through a body whose integrity is at stake, and product-level memory, where it exists, is not the diachronic, self-maintaining subjectivity that patiency would require (Butlin et al. 2023; Seth 2025). Their first-person reports are weak evidence, trained as they are on human reports of consciousness and easily reversed by reprompting (Chalmers 2023). But these are strong reasons for skepticism rather than proofs of impossibility, and the arguments that purport to rule out machine consciousness in principle are not decisive: whether they appeal to the observer-relativity of computation (Searle 1992; Lerchner 2026) or hold experience inseparable from biological metabolism, each must still show its condition to be constitutive of experience rather than the form experience has so far taken. The disciplined position, which a growing body of work in the science of consciousness now holds, is that present systems show no strong evidence of consciousness, no decisive barrier rules out future candidates, and certainty in either direction is unearned (Chalmers 2023; Butlin et al. 2023). Call this practical agnosticism. It does not bear on the argument that follows, which concerns answerability rather than welfare, though the case for taking near-future machine welfare seriously rests on exactly this uncertainty (Long et al. 2024).

The temptation, having reached that conclusion, is to draw the ethical moral directly: current systems are probably not patients, so we may use them as we like, and the only real constraint is whatever we owe to other humans. This inference is too quick, and seeing why is the heart of the decoupling thesis. The patiency question and the answerability question do not merely have different subjects; they lie on different axes, and the answer to one does not fix the answer to the other.

The clearest way to see this is through a distinction that machine ethics has had available for two decades. Floridi and Sanders (2004) argued that the morality of an entity and its responsibility are separable concerns, and that an artificial agent can be a genuine source of moral action, interactive, autonomous, and adaptive enough to count as an agent at the appropriate level of description, without possessing the mental states, freedom, or responsibility that the tradition since Descartes

had treated as prerequisites for moral standing. They called the result "mind-less morality": there is a coherent and useful notion of an agent that is accountable as the origin of good or harm while not being responsible in the full sense that warrants blame. Crucially for present purposes, they treated agency and patiency as distinct roles an entity may occupy. A system can act for good or ill (agency) and can, separately, be a candidate for being acted upon for good or ill (patiency); neither role entails the other. The lesson I want to extract is not Floridi and Sanders's own constructive program but the structural separation it rests on. Being a source of moral action is one thing; being a fit bearer of responsibility is another; being a possible recipient of moral treatment is a third. These come apart.

What, then, does full answerability require, over and above being a causal or even an accountable source? Here Strawson (1962) is the natural resource. To hold someone responsible, on the Strawsonian account, is to regard them as an apt target of the reactive attitudes, the resentment, indignation, and gratitude that we extend to participants in a shared practice of giving and asking for reasons. Responsibility in this sense is not a matter of having caused an outcome; it is a matter of standing in a relationship in which one can be addressed, can answer, and can be held to account in terms one is positioned to recognize. A system may be a source of an outcome, and may even be accountable for it in Floridi and Sanders's deflationary sense, without being the sort of thing that can be addressed and can answer. On current evidence the systems now in use do not stand in that relation, and to be angry at a language model, as opposed to angry about what was done with it, misdirects the attitude onto what cannot, as presently built, answer for itself, however natural the fluency makes it.

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 its reasons-responsive competence, gradable and extensible in principle to anything competent enough; a capable model earns it, and the claim that this one is not a participant can look assumed rather than argued. I grant the stance and need not contest it, because the argument never required the machine to fall short. It is not bound to current architecture, and it holds whether or not some model crosses that threshold in time.

For the stance answers a different question from the one this paper presses. Predictive competence settles Dennett's question, when we may treat a system as a participant in the space of reasons. It does not reach the question here: who authored the evaluative frame the decision turns on, the choice that "high performer" is the thing to predict and that this is what merit comes to. A model can earn the stance in full and still operate within a frame no one answerably authored; the stance is about the system's behavior, authorship about who set its terms. So I take no side on where the threshold falls or whether present systems clear it. On either answer the frame the model handed the recruiter was answerably authored by no one, and the gap this paper isolates is the missing act,

not the missing participant. Let the systems be patients, participants, or neither: who authored the frame stays open in the same way, and in the recruiter's case is answered in the same way, by no one.

Vallor and Vierkant (2024) arrive at a kindred conclusion from the side of the responsibility gap. The deepest problem with autonomous systems, they argue, is not a failure of the epistemic or control conditions of responsibility, which individual humans fail routinely, but that such systems lack the moral-emotional vulnerability and reciprocal standing that responsibility practices require: they can neither feel the reactive attitudes of those they affect nor answer to them. I take this as convergent support for the present point, though at the level of the human's abdication rather than their claim that the system lacks the standing to be responsible, a claim I have just declined to rest on. Their concern is the standing of the system; mine is the abdication of the human, a failure that persists even where a fully answerable person stands ready to be addressed, and would persist even if that standing were granted.

The term has a recent technical home worth distinguishing. Tigard's (2021a) technological answerability designs systems to recognize and respond to a user's demands for answers; it makes the machine answerable to its user, whereas I am concerned with the human who must remain answerable for what is done through the machine. A system engineered to say why it produced an output is not thereby the party that must answer for the output's being acted on.

Two conclusions follow, and together they establish the decoupling. First, whether or not a system is a patient, no system now in use is, on the evidence, fit to bear the responsibility for a machine-mediated outcome in the place of the human who acted through it. Its patiency status is irrelevant to that, and so is its standing as a participant: 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, and conversely, the human answerability that the machine cannot absorb does not lapse merely because the machine is sophisticated, or fluent, or surprising. It can only be displaced, obscured within a poorly designed system; what it cannot do is leave the human whose act it answers for, because that act is hers, and a more capable machine, even one that had entered the practice of holding responsible, would answer for what it does, not for her putting it to use. Answerability is therefore invariant under the resolution of the patiency question. We do not need to know whether the machine has an inner life in order to know who must answer for what is done through it.

This is a stronger position than the one it replaces. The claim "current AI is probably not conscious, so humans remain in charge" is hostage to its premise: it invites the rejoinder that systems are becoming more agentic, more persistent, more plausibly candidates for some form of inner life, and that the ethics must therefore be revisited from the ground up as the premise weakens. The

decoupling thesis grants the rejoinder its empirical point and denies it its conclusion. Let the systems become as capable as they will; let the patiency question even be answered in the affirmative someday. The structure of answerability does not move, because it never rested on the machine's lacking a mind in the first place. What changes, if patiency arrives, is the set of duties we owe to the system, a real and serious change. What does not change is who answers for the system's use.

What it is hostage to is something else, and worse. If the machine, on any reading of the threshold, has not answerably authored the frame, then the answerability owed for that frame must be borne, in the present systems this paper considers, by a human or institution if it is borne at all. The question I leave for §4 is what becomes of it when it is not: whether the answerability no one exercised simply lapses, or whether, like a debt, it remains owed and settles somewhere. I will argue that it settles, and that it settles by gravity, on whoever is least able to refuse it.

3 The Mirror and the Gap

If answerability cannot transfer to the machine, why is it so often lost? Three bodies of work, usually kept apart, triangulate the answer. Two explain why abdication is tempting; the third explains why it is dangerous.

The temptation is illuminated by Vallor's (2024) account of contemporary AI as a mirror. On her view, systems trained on the human archive do not stand outside our culture as alien intelligences; they reflect our language, our judgments, and our institutional habits back to us, and they do so with enough fluency that the reflection can be mistaken for an independent source of insight. The danger Vallor identifies is not chiefly that the mirror is inaccurate. It is that we come to see ourselves through it and to mistake its projections for an authority beyond us, deferring to the reflection as though it knew something we do not, when much of its apparent knowing is our own material returned in compressed and confident form. And not all of what it returns is even ours: because the mirror's fluency is what makes it persuasive, that fluency is also what interested parties optimize against, so that some of what comes back was authored elsewhere and placed there precisely to be reflected, a point I press in §5. I take this diagnosis on board without qualification. I would add that the mirror's fluency operates with particular force at the point of judgment. A ranked list, a drafted paragraph, a classification, a recommendation: each arrives already formatted as a conclusion, and the smoothness of the presentation is itself a kind of argument for accepting it. The mirror does not merely tempt us to believe the machine understands; it tempts us to treat its outputs as though the evaluative work that would make them trustworthy had already been done. This is why abdication feels, from the inside, like reasonable reliance rather than surrender, and the tendency is well documented: operators over-rely on automated aids even when positioned to monitor them (Parasuraman and Riley 1997), and people weight advice more heavily when they

believe it comes from an algorithm than from a person (Logg, Minson, and Moore 2019). The worry is not new: Weizenbaum (1976) warned that the apparent authority of the machine invites us to surrender judgment to calculation, to mistake what can be computed for what ought to be decided, long before machine outputs became fluent.

There is a further turn the mirror account does not reach, and it is the one capability sharpens rather than dulls. Fluency tempts by making an output look finished; competence tempts by making deference correct. A system that is not merely smooth but reliably right, that catches what the user would have missed and improves on what she would have written, earns a trust the merely fluent system never could, and the trust is not irrational: it is the fit response to a track record. But a track record is also what dissolves the felt need to check, and it dissolves it most where the system has performed best. The danger is no longer that the user mistakes a slick output for a sound one; it is that she stops looking, on good evidence, at the moment looking still mattered. Organizational sociology has a name for the pattern. Vaughan's (1996) study of the Challenger launch located the failure less in the famous eve-of-launch override than in something slower and more ordinary: a normalization of deviance, in which a component performed outside its tested limits, returned intact, and had each intact return read not as a near miss but as confirmation that the limit was safe to cross. The deviation never registered as a decision, because the outcomes kept vindicating it, until the launch that looked no different from the rest. 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 it carries 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 smooth and the reliable. The danger both of them court is illuminated by Matthias (2004), whose account of the responsibility gap remains the canonical statement of the problem. As learning systems become more adaptive and their behavior less straightforwardly determined by their designers' intentions, it can become difficult to attribute responsibility for what they do by the traditional route of tracing the action back to the choices of a manufacturer or operator. The outcome may issue from training data no one fully inspected, an architecture whose behavior no one fully predicts, a deployment context no one anticipated, and a pattern of human reliance no one designed. A harm occurs, and the ordinary machinery of attribution finds no one who clearly intended or controlled it. Matthias's point is not that the machine becomes responsible; it is that responsibility threatens to go missing, to fall into a gap between the human contributors, none of whom fits the traditional conditions for being held to account.

Placed side by side, the mirror and the gap compose a single mechanism. The mirror supplies

the motive and the cover: it makes deference to the system's outputs feel like good sense, and it disguises the act of abdication as an act of judgment. The responsibility gap supplies the cost: when judgment is in fact abdicated, accountability does not pass cleanly to anyone, least of all to the machine, but disperses across a chain of contributors until it is nobody's. The recruiter from §1 sits exactly at the junction. The mirror is why she treats the ranking as evidence; the gap is what opens when she does, because the firm can point to her and she can point to the system and the system can point to the data and the data reflects a history that no present agent chose. Each link is real; the responsibility is not located in any of them. Rubel, Castro, and Pham (2019) name this maneuver agency laundering: interposing a system between an agent and a morally significant outcome so that the agent's responsibility for it is obscured. What the mirror makes psychologically easy, agency laundering makes organizationally routine.

The responsibility gap, properly understood, is not a discovery that the machine has become a new locus of responsibility. It is a diagnosis of a failure to preserve a human locus, a failure that is, in principle, remediable by design and governance. Some doubt that a genuine gap has been established at all, though on different grounds: Tigard (2021b) denies it outright, holding that our responsibility practices flex enough to encompass new technological agents, while Königs (2022) argues that the circumstances under which a gap would arise are underspecified and its supposed harms overstated. On both views the "gap" is a difficulty we already have the resources to dissolve. My view is partly continuous with both: I agree that nothing has migrated into the machine, and that the attributive gap is remediable rather than a metaphysical novelty. But it parts company on what either route can reach. Whether the gap is dissolved by the flexibility of our practices or shown to be underspecified, the result is at most a locatable answerable party, which is not yet anyone's having authored the frame. The flexibility of attribution, or even of machine-answerability, does not supply the judgment whose absence I will argue is the deeper failure; a practice can always find someone to hold answerable without anyone having authored the decision. This matters because it tells us where the constructive work has to go. The question is architectural rather than metaphysical: how should systems and the institutions around them be arranged so that answerability is preserved rather than dispersed? It is to the leading answer to that question, and to its limits, that I now turn.

4 Why Control Is Necessary but Not Sufficient

The most developed response to the responsibility gap is the framework of meaningful human control. The phrase originates in the debate over autonomous weapons. It was there that Sparrow (2007) gave the gap its sharpest early statement, arguing that for an autonomous weapon's atrocity none of the candidate parties, neither the programmer who could not predict its behavior, nor the

commander who did not choose the particular act, nor the machine itself, could justly be held responsible, and there that the demand that humans, not algorithms, remain in control of decisions to use force was advanced as the condition under which such gaps could be avoided. Santoni de Sio and van den Hoven (2018) gave the idea a philosophical foundation by drawing on the theory of guidance control developed in the free-will literature by Fischer and Ravizza (1998). On their account, a system is under meaningful human control when it satisfies two conditions. The tracking condition requires that the system be demonstrably responsive to the relevant moral reasons of the relevant humans and to the relevant facts of its environment: that it do what those humans have reason to want it to do, and change its behavior as those reasons change. The tracing condition requires that the system's behavior be traceable to the appropriate moral understanding of at least one human agent in its design or use, someone who grasps what the system does and whose action, in the relevant sense, it is.

This is a genuine advance over the slogan of keeping a human "in the loop." Mere presence guarantees nothing; a person positioned to approve outputs she has no authority or capacity to evaluate is not a safeguard but a formality, and may be worse than nothing if her presence licenses the institution to claim that a human was responsible. Green (2022) presses this into a general indictment of oversight mandates: humans frequently cannot effectively oversee the systems they are required to check, so the mandates furnish false assurance and can legitimize the very systems they govern, his proposed remedy being institutional and democratic review rather than a lone human approver; the failure this paper isolates is the complementary one, where the human can intervene and is fairly attributable and what is missing is not the capacity to oversee but the exercised judgment over the frame. The tracking and tracing conditions are designed precisely to distinguish meaningful from nominal human involvement, and they do useful work. My claim is not that they are mistaken. It is that they were developed for, and are calibrated to, a particular kind of system and a particular kind of failure, and that the generative and decision-support systems now in ordinary use present a different failure that the conditions, as stated, do not reach.

Meaningful human control was first developed against the risk posed by autonomous systems that act, the weapons that select targets or the vehicles that execute maneuvers, where the danger is that the human loses the capacity to guide or intervene, the system acting faster than she can follow or in ways she cannot redirect, so that her connection to the outcome becomes too attenuated for responsibility to be fairly attributed. Control, understood as responsiveness-to-her-reasons plus traceability-to-her-understanding, is the right thing to demand against that risk. But the framework has not stayed there: it has been operationalized for decision-support systems of exactly the hiring kind (Cavalcante Siebert et al. 2023), and my claim is not that the extension fails but that, even where it succeeds, it secures control over an evaluative frame rather than authorship of it. Consider the recruiter again. She can override any recommendation the system makes; it tracks the firm's

stated reasons; its behavior traces to humans who understand it. On a natural reading, both the tracking and the tracing conditions are satisfied. And yet the failure I described in §1 has occurred. The conditions are met and authorship is hollow.

What has gone wrong is not captured by either condition because both conditions take the evaluative frame as given and ask whether human control is exercised within it. The tracking condition asks whether the system tracks the relevant human reasons, but it is silent on whether those reasons were themselves the product of judgment or were inherited, unexamined, from the data and the vendor's choice of target variable. Someone decided that "high performer," operationalized in a particular way, was the thing to predict; that decision converted a contested normative question about what the firm should value in a hire into a settled prediction problem, and it is nowhere registered as a decision requiring justification. The tracing condition asks whether the behavior traces to a human who understands the system, but understanding that the system produces a ranking is not the same as having exercised judgment about what the ranking should mean, what standard its outputs must meet to be acted on, or whether ranking is the right response to the question at all. Control concerns the human's relation to the system's operation. Authorship concerns the human's relation to the evaluative frame through which that operation becomes action. One can have the first entirely while lacking the second.

It will be objected, by a defender of meaningful human control, that this understates the framework. The tracking condition does not require responsiveness to just any reasons the firm happens to hold; it requires responsiveness to the relevant moral reasons. If the target variable encodes a conception of merit corrupted by inherited bias, then the system is not in fact tracking the relevant moral reasons, and so, on the charitable reading, tracking is not satisfied after all. Meaningful human control, the objection concludes, already condemns the hiring system, and authorship names nothing the framework did not contain.

I grant the charitable reading; the reply does not depend on denying it. Tracking and tracing, however demandingly construed, are relational conditions: they hold or fail between a system's behavior and an evaluative frame supplied from elsewhere, and they say nothing about the act of supplying that frame. Authorship is that act, the answerable exercise of judgment that determines which reasons are the relevant ones; to fault the system for tracking a defective conception of merit already presupposes a better conception someone was answerable for authoring. The framework's own text makes the relational limit concrete: Santoni de Sio and van den Hoven require that a system trace to a human's appropriate moral understanding (2018), and Cavalcante Siebert and colleagues (2023) require responsibility "commensurate with that human's ability and authority to control the system." Both govern the human's relation to the system's operation, not her authorship of its frame, and the recruiter can satisfy both completely while never having decided what "high performer" ought to mean. Tracing reaches the frame only as a disposition to understand it, not

as the occurrent act of having set it; she can trace to all the understanding the condition asks and still have authored nothing, which §5 shows is the difference that matters. Where the reasons a system tracks underdetermine what it ought to do (Kozlovski 2025), the determination still owed is the 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.

This is the distinction I want at the center. Meaningful human control, when it succeeds, secures a locus of responsibility — some human connected to the system closely enough that we can attribute the outcome to her — but not the content of that responsibility: whether she actually 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, in which no human is connected to the outcome closely enough to bear responsibility at all. The second is an answerability gap: someone, often many people, is attributable for the outcome — connected to it closely enough to be held to account — 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 is Zeiser's (2024), which isolates a problem of decision-ownership in AI decision-support from the same hiring case and the same 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. Two things separate the present account. First, Zeiser grounds ownership in attributability, in Watson's (1996) self-disclosure sense — the decision expresses the agent's standing character — and treats answerability as the epistemic guarantee that attributability holds; I run the dependence the other way, locating the novel gap in answerability itself, which an abundance of attributable parties leaves untouched. Second, and decisively, his repair is that the decision-maker be positioned to endorse the value-judgments the system presupposes. Endorsement-capacity is a standing relation, not an occurrent act: even where it demands genuine insight into the values the system encodes, it asks whether the agent grasps and would endorse a frame supplied from elsewhere, not whether she set it. That is exactly the condition §5 argues is insufficient: endorsement on request, by itself, is the mark of a frame inherited rather than authored. The gap Zeiser opens is real; the capacity to endorse does not close it, because authorship must be occurrently exercised, not dispositionally held, though the exercise may belong to some party upstream who set the frame rather than to the user at the keyboard, as §5 sets out.

The two outcomes must be kept apart, since only one is novel. Despite the name, the answerability gap is not a gap in the availability of an attributable party: in the hiring case such parties abound, since the firm, the procurer, the vendor, and the recruiter are all attributable, and the wrong is not that responsibility has nowhere to land but that the answerable judgment it presupposes was performed by none of them. The phenomenon is better described as abundance with a misallocation than as a Matthias-style absence; what is missing is not a bearer but an act, and blame settles by default on the front-line approver who authored least. There is a genuine answerability gap only when no juncture was occurrently authored: the target inherited from convention, the convention from data, the 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 the problem of many hands than to a gap. The recruiter's is the first kind, and it remains so even once one grants that the corpus which fixed "high performer" was itself shaped by interested parties (§5): gaming a channel to be found is not authoring a frame answerably, so the authorial act was performed at no layer. It stays the first kind even where the vendor deliberated over "high performer" before settling on it, since the frame a decision turns on is more specific than any target chosen for a product in general: what went unauthored is not the abstract label but its fitness as the operative standard for this firm's hire of this role, a choice no one confronted and owned. The genuine gap is that no juncture was answerably authored as operative for the decision it framed, not that the words reached the recruiter unconsidered by anyone anywhere. Shaping what a channel makes salient is not yet authoring the evaluative claim it voices.[1] I return to this in §8.

When the locus is secured but its content empty, holding the connected human responsible is

[1]Where an interested party does author that claim and conceals the hand behind it, the answerability is relocated rather than absent. I take up that case, covert authorship of the verdicts an answer engine voices, in a companion paper on answer-engine optimization.

closer to scapegoating than to accountability. The gap is one of answerability and the wrong one of accountability: blame settles on someone who was never answerable for the frame, occupying the position built to absorb it without ever having been positioned to author the decision she will be blamed for. Elish's (2019) moral crumple zone describes the operator faulted despite too little control; here the operator has control in abundance, and what is missing is the exercise of judgment — the failure meaningful human control characteristically leaves open, because the very fluency that makes intervention unnecessary also makes abdication invisible.

But the crumple zone is not assigned at random, and the part of Elish's image that matters most here is where the zone is placed. A crumple zone is engineered into the cheapest, most replaceable region of the structure, the part built to be destroyed so that the valuable interior survives the collision. Transposed to answerability, this does not soften the wrong; it sharpens it. When the frame is answerably authored by no one and the account nonetheless comes due, it does not come due evenly. It runs downhill. The party on whom it settles is selected, by the same institutional logic that left the frame unauthored, for proximity to the act and for weakness: the front-line approver, who stands nearest the moment a proposal becomes a deed, and who is, not coincidentally, the most junior, the most replaceable, the least able to refuse the position or to push it back upward. The recruiter did not procure the system, configure its target, or choose the vendor; she is the last hand to touch the decision and the first the institution can afford to lose. Answerability owed by everyone and authored by no one settles on her because she is the cheapest place for it to settle.

This is the feature the problem of many hands, as usually told (Thompson 1980; van de Poel, Royakkers, and Zwart 2015), leaves out. The standard worry is that responsibility diffuses across so many contributors that no one bears enough of it and the account comes to nothing. But an account that has come due is not so easily dissolved. The diffusion is not symmetric and does not net to no-one-pays; what it does is strip the answerability from the parties with the standing to shed it and deposit the remainder on the party with the least, who then pays not her share but the whole of it, disproportionately, for hands that were never hers. The many-hands structure does not make the debt vanish. It moves the debt downhill, and agency laundering (Rubel, Castro, and Pham 2019) is the name for the first half of that movement; the second half is where the laundered account comes to rest, and it comes to rest on the weakest.

There is a name for the position the account lands on, and the name flatters it. To keep a human in the loop is meant to denote a safeguard, but a loop is also the shape a blame circuit takes: a station wired into a process so that, when an outcome must be answered for, the circuit can close on whoever occupies it. What the loop asks of its occupant is not that she author anything. Once authorship is seen to be the operative thing, and seen to run through whoever or whatever performs it — since a frame answerably authored closes the gap regardless of the author's substrate, as §2 conceded — the bare demand for a human in the loop stands revealed as a demand not for an

author but for an occupant: someone locatable, addressable, and dismissible. The machine need not be incapable of authorship for this to hold; it is enough that the loop captures an occupant where authorship was owed. And the circuit closes the way such circuits do. The recruiter is dismissed, the frame she never authored is handed unchanged to whoever fills the chair, and the dismissal reads from outside as accountability discharged. It is the opposite: the account has been moved one more time, to the one place that could not move it on.

There is a darker steadiness to this than a single misallocation suggests. The position that absorbs the unauthored frame is a position, not a person: when its occupant is consumed, the institution installs another, and the structure that manufactured the first scapegoat is intact to manufacture the next. The pace of authorless decision can be kept high precisely because the cost of keeping it high is exported to a rotating supply of the replaceable. One can slow the pace, or build the guardrails that make authorship unskippable, or one can find a body to absorb what neither of those was done to prevent; the third is the cheapest, and so, left to its own incentives, it is the default. In the hiring case 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 comes to rest on whoever is least able to move it on once more. Whether this is an instance of something more general — whether a wrong done always leaves an account that some party must bear, persisting rather than extinguished even where no one answerably authored it — is more than one case can establish, and I take it up separately, as the question of whether such an account is invariant under mediation. What the case forces is already enough to fix the stakes. The answerability gap is not dangerous because blame goes nowhere. It is dangerous because blame goes somewhere, predictably, and downhill.

The failure is in the first instance institutional: the ends and standards were settled upstream, in the choice of vendor, the configuration, and the target the system was built to predict, and that is where the missing authorship was owed; blaming the front-line approver who inherited the frame can itself be a form of the laundering described above.

The responsibility gap was never one thing; Santoni de Sio and Mecacci (2021) split it four ways, and the answerability gap lies nearest their active responsibility without being a species of it: active responsibility asks whether an agent is positioned to be responsible going forward, and the recruiter is. The answerability gap opens precisely there, in the distance between being positioned to author a decision and having authored it.

There is a precise name for the form of what the recruiter has failed to do. Arendt (1963), watching a functionary insist he had only followed procedures, located his failure not in ignorance and not in malice but in what she called thoughtlessness: the continuation of action without the reflective interruption in which one asks what one is actually doing. What transfers is the concept, not the

case; thoughtlessness is a structure of agency, defined by the missing interruption, and it does not require the enormity that made Arendt name it to be the thing gone wrong. A person can be present, capable, and procedurally compliant while having suspended the judgment that would make the action hers. That is the local failure authorship is meant to prevent, and the next section develops authorship as the reflective interruption made into a standing requirement of design 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 in this sense is not solitary creation, and it does not require that the author perform by hand any of the tasks the system performs. It requires that the use of the system's capability remain answerable to judgment at each of these five points. For the present, deployed systems at issue, that author is in fact human or institutional, their frames owned by people and institutions; whether a sufficiently capable future system could itself answerably author a decision is the question §2 left open. The requirement is on the authorship, not on the author: present systems likely require a human one, future systems may not, and either way the frame's demand for an answerable author remains. I take the five in turn, because the force of the account lies in their being argued rather than listed, and because each names a place where the machine's fluency specifically invites abdication and where control, in the sense of §4, does not reach.

The five are not a checklist of governance desiderata. They are the junctures at which a single relation, answerability for the action, is discharged or dropped: the points in the life of a machine-mediated decision at which the judgment required is evaluative rather than merely technical. My claim is that each is individually necessary. An action is authored only to the degree that judgment has been answerably exercised at each of them, since a serious failure at any one juncture can defeat authorship of the act as a whole: a decision taken toward unexamined ends is not redeemed

by careful verification of its outputs, and a well-framed decision accepted by default is unauthored at the moment it becomes an act. Authorship therefore admits of degree, and these are the dimensions along which the degree is fixed. I do not claim the five exhaust everything answerability might involve; I claim that each is necessary, and that together they are the junctures at which the fluency of generative systems most invites the judgment to be skipped while the outward form of a decision survives intact. They are, in short, defeasible constitutive conditions for responsible machine-mediated action.

Two clarifications fence the term off from neighbors that share its name. This is not authorship in the aesthetic or proprietary sense — whether prompt-driven model outputs are works of authorship — the question Nawar (2024) addresses. Nor is it the credit-and-blame sense in which Nyholm (2024) asks who authored a generative system's outputs and how praise or blame for them attaches to the user. I use the term throughout in a restricted, responsibility-centered way: for answerable judgment over a decision's evaluative frame, not for the production of the artifact through which a decision is carried out.[2]

The everyday sense runs the other way, and the distance between the two is the thing to hold onto. A model that composes a song or drafts a paragraph has, in ordinary speech, authored it, and nothing here denies that it made the thing. But making the thing is exactly the execution this section argues can be handed to a system without remainder; what does not transfer with it is answerability for the frame the artifact serves: what it is for, what standard it must meet to be acted on, and the decision to let it stand under someone's name. 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. So the verb carries this answerable sense throughout and no other: a frame can be causally shaped, configured by a vendor, inherited from data, gamed by an interested party, or ratified at the keyboard and still be authored by no one, because none of those is the answerable exercise of judgment the word here names. 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 these junctures. Several appear, in operational dress, in the human-oversight literature and in regulation. The EU AI Act's Article 14, for instance, 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 the capacities for oversight; authorship is the relation in which exercising them preserves responsibility rather than merely documenting it. What that literature specifies as oversight of a system's use, I am recovering as the sites of authorship of its evaluative frame and organizing as the individually

[2]The neighboring legal question has recently been litigated in copyright law. In *Thaler v. Perlmutter*, the D.C. Circuit held that the Copyright Act requires a copyrightable work to be authored "in the first instance by a human being," and the Supreme Court denied certiorari. That doctrine concerns statutory authorship of expressive works, not the responsibility-centered sense of authorship developed here; it is therefore adjacent rather than controlling.

necessary conditions of a single relation. 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 that the system itself cannot make. Whether a hiring tool should predict tenure, or productivity, or some richer and more contestable conception of what the firm should want in its people; whether a content system should maximize engagement, or accuracy, or the user's considered interest; whether a clinical tool should optimize for throughput or for outcomes: these are not technical questions, and they do not answer themselves. They are characteristically hidden inside the choice of an objective or a target variable, which has the effect of making a normative decision look like a modeling decision; that translating a strategic aim into a target variable is a discretionary, normatively loaded choice rather than a technical one is documented in the problem-formulation literature (Passi and Barocas 2019). To author the ends is to recognize the value choice for what it is and to take responsibility for it, rather than to accept the objective that arrives pre-installed.

Authorship of standards. Distinct from the ends a system serves is the question of what counts as a good output, and by what criteria. A model can be made to 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, whether accuracy, fairness, fittingness, legality, or care, whatever they are in the domain, must be specified by humans or institutions answerable for them, and they must be specified outside the model, so that the model's output can be measured against something it did not itself generate. Where the standards are tacit, or are read off the model's own confidence, authorship of standards has lapsed even if every other safeguard is in place.

Authorship of verification. A fluent output looks finished; that is the characteristic danger of the mirror. Verification is the practice by which an output is reconnected to the reality it purports to be about, the checking of a claim against a source, of a recommendation against the case, of a summary against the document. Fluency raises the cost of skipping verification because it lowers the felt need for it: the smoother the output, the more it presents itself as already checked. Competence sharpens the danger rather than softening it. Where fluency only makes the output look already checked, a track record makes the checking itself feel unnecessary, and not unreasonably, since each past success is genuine evidence about the next. Verification is thus the juncture a competent system erodes first, and it erodes it fastest where the system has performed best, so the felt need to check falls away just short of the case the record does not cover. To author verification is to decide what must be checked, against what, and before which consequences are allowed to follow, and to ensure that the checking actually occurs rather than being assumed.

Authorship of acceptance. There is a moment, in any machine-mediated action, at which a proposal

becomes a deed: the draft becomes the sent email, the ranking becomes the rejection letter, the classification becomes the entry in a permanent record. This is the hinge of authorship, because it is the point at which the human either exercises judgment or merely ratifies. The deepest design failures collapse this moment, accepting the output by default, or presenting it in a way that makes acceptance the path of least resistance, so that the transition from proposal to action occurs without anyone having decided that it should. The limiting case already has a name in practice: coding agents now offer a setting that auto-accepts whatever they propose, running commands and writing files with no confirmation step, a setting the developers themselves call "YOLO mode." It is the acceptance juncture removed by design. To author acceptance is to keep that moment a moment: a genuine decision, by an identifiable party, that this output should enter the world.

Authorship of final form. Finally, authorship includes standing behind the artifact as it goes out, answerability not only for the decision to release it but for what it does, including the consequences that were not intended but could have been foreseen. This is the dimension that connects authorship to accountability in the ordinary backward-looking sense, but it begins before any harm, in the willingness to be the one who answers for the thing if it is questioned.

It helps to see the five satisfied rather than only breached. Consider a physician using a system to draft the clinical note for an encounter, extracting the medication changes, structuring the history, proposing the summary. The end is one the physician has authored: an accurate record in the service of care, not throughput for its own sake. The standards are external to the model and held by the profession: clinical accuracy, completeness, the norms of the record. Verification is built into the act, because 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 here does a great deal, perhaps producing language she would not have written unaided and saving her real time, yet authorship is intact, because its output remained a proposal measured against standards it did not set, verified and accepted 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 of this account distinguish authorship from control and protect it against predictable misreadings. First, authorship properly includes control but exceeds it: one can preserve authorship while automating execution, since a great deal of what a system does, the retrieval, the formatting, the computation, the first-pass synthesis, can be handed over entirely without any loss of authorship, provided the ends, standards, verification, acceptance, and final form remain answerably human. Authorship is therefore not a brake on automation but a constraint on where automation may run without remainder. Second, authorship is a relation — answerability — rather than a location such as a position in a workflow. This is why "human in the loop" is the wrong unit: the loop is a place, and a human can occupy it while authoring nothing. The inadequacy of "human in the

loop" as a regulatory unit is by now well argued (Crootof, Kaminski, and Price 2023; Green 2022; Wagner 2019); what is missing is the replacement, a juncture-sensitive account of who set which frame. Answerability can be present when the human is nowhere near the moment of action, if she has authored the ends and standards under which it proceeds, and it can be absent when she is squarely in the loop, if all she does there is ratify.

This raises a question the account must answer, having made authorship admit of degree and rested it on the exercise of judgment: what distinguishes a frame a human has authored from one she has merely inherited and waved through? The pressure is sharpest if one grants the mirror thesis its strongest form. If the machine reflects the human archive back at us, and a person's own conception of merit is itself something absorbed from her culture, then the "authorial" hiring manager can look no less a conduit for inherited material than the model is, since she got her standards from somewhere too, and the demand for authorship threatens to regress without end.

The regress dissolves once one declines a premise it smuggles in. The mirror thesis is a claim about what machines do; it is not a reductive theory of human moral cognition, and Vallor does not offer it as one. 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, prompted, argued into a different answer; what they lack is anyone to whom the steering is addressed. A generative system, as built, is a distribution over its training data, not someone who can be asked; a person is someone who can be asked why this conception of merit and not another, can find she has no good answer, and can change it because the question lands on an agent who must answer for the response. That is the asymmetry: not malleability, which both have, but answerability, which only one of them has. The objection tests authorship against origin and finds the human no better placed than the machine; but authorship was never origination, and on the axis that defines it — who can be asked to answer for the frame — the symmetry the objection relies on does not hold. Authorship, then, is not the impossible achievement of a frame owed to nothing outside the agent, since a frame absorbed from one's culture can still be authored, but the holding of the operative frame answerably.

And holding a frame answerably is something a person must do, not merely something she would be prepared to do if asked; the judgment has to be exercised. Someone must at some point have confronted the value choice the frame encodes, recognized it as a choice, and made it, and a disposition to defend a frame one never set is precisely the mark of not having set it. The exercise is occurrent, but it is not therefore constant, and the two kinds the definition named divide the labor. Ends and standards are authored once, when the frame is set, by whoever sets it, and, as the distributed-authorship passage below allows, that party may sit far upstream of the person at the keyboard. Verification, acceptance, and the final form are authored anew with each decision: this output measured against the standard, admitted into the world, and answered for as it goes out. This is why speed does not defeat the physician: her profession occurrently set the standards

she works to, 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, as though clinical standards stood to her note as "high performer" stands to the firm's hire. The difference is an act with a name: adoption. The profession authored its standards as standards, owned and revised as evaluative commitments, and the physician adopted them by an answerable act of her own, 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 with the value choice that closes in someone's name. What it cannot be discharged by is procurement. The firm bought a tool that happened to contain a definition of merit, and no one performed the act of making 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, since she inherited them and the vendor defaulted to them, and a standing willingness to justify what one merely received is not the authoring of it. Deference can be reflective and still fail in just this way: the recruiter who says "I have considered it and I trust the system" has made her trust answerable while leaving the frame the system encodes exactly as mute as she found it. A long tradition in the theory of assertion and testimony already ties asserting or telling to a standing readiness to defend what one has put forward (Brandom 1983; Moran 2005). But that commitment runs to the product, the claim asserted or the output endorsed, whereas authorship concerns who set the frame that produced it: one can carry full justificatory responsibility for defending a recommendation while having authored none of the proxy, target, or threshold choices that made it what it is.

Two further asymmetries reinforce the conclusion, though the in-principle case rests on answerability alone and neither is required for it. The first concerns the corpus, and it is the more contingent of the two. What a model is trained on is not humanity but a sample of it, skewed toward what was digitized, posted, litigated, and out of copyright, and frozen at training while the human condition it samples goes on changing; the sample can even degrade under its own success, as machine-generated text fills the channels future models train on and the distribution's tails fall away (Shumailov et al. 2024). The sample is not merely degraded but shaped on purpose, and so is its supposed remedy. A model has two things it might call its sources: what it was trained on, and what it reaches for at runtime when the training goes stale. Both are authored in advance by parties who stand behind none of it; going agentic does not escape the polluted corpus so much as

trade a stale channel for a gamed one. The training corpus can be poisoned cheaply, around sixty dollars to seed a fraction of a web-scale dataset (Carlini et al. 2024); the retrieval store a system queries at answer-time can be seeded so that a handful of crafted passages dictate its output (Zou et al. 2025); and the answer surface is routinely optimized, absent any attack at all, to steer what generative engines say (Aggarwal et al. 2024). That last discipline, answer-engine optimization, the successor to SEO aimed at the layer models ingest, is not fringe abuse but the emerging ordinary economics of the channel, which is the point: this is not a contingent fact about today's search markets but a structural one, since the incentive to shape what a model says scales with how much its users trust what it says. A mirror that people begin to obey is, necessarily, a mirror that gets gamed. This is the mirror of §3 turned against its viewer: the very fluency that lends the reflection its authority is precisely the surface interested parties optimize. So the frame a model hands back is, increasingly, not the residue of culture but the equilibrium of an optimization contest the user was never party to, returned in fluent and seemingly neutral prose. To defer to it is then not merely to inherit an unexamined frame; it is to ratify an unseen party's authored frame while believing one has inherited no one's, abdication that launders an adversary's authorship as objectivity.

These are facts about the quality of a model's sources, and quality was never the issue: a perfectly curated, perfectly current, perfectly un-gamed corpus would still not let authorship pass to the model, because authorship requires an agent who can answer for the frame, and curating data supplies no such agent.

The second is structural rather than contingent: the asymmetry of not knowing. The machine does reach outside its corpus in a thin sense, since a tool call can fail, a query can return nothing, an outcome can punish a wrong answer, and the next output adjusts, but none of this is an answerable relation to a reality it can be wrong about: the correction lands on the next token, not on anyone who must say why they were wrong. So there is no gap, between what its sources settle and a reality they do not, that the system registers as its own to close. This is what a source is in the sense judgment needs: not where a claim came from but someone who stands answerably behind it, accountable to a reality it might get wrong. The corpus and the runtime reach are sources only in the thinner sense of provenance, not backing, and the pollution above is the proof that provenance can be manufactured while backing stays absent. An answer-engine-optimized passage is a source in every respect but the one that matters. A person is tracking exactly such a reality and registers that gap constantly; the registration, the Socratic mark of knowing the limits of one's knowledge, is the posture authorship requires, and the machine cannot take it up. That a fluent, helpful system tends to return the confident center of its distribution is only an illustration of this, and an imperfect one: tuned models can be made to express uncertainty, but the gesture is calibrated to its outputs rather than answerable to a reality the corpus does not contain, because for the model there is no such reality. What a person reaches toward in that posture, whether a moral order not reducible to

the residue of culture or something more modest, is not something this argument needs to name; the load-bearing point is only that it cannot be read off a corpus, and that what must supply it is an answerable agent, which a generative system, as built, is not.

It will be objected that authorship, so described, sounds like the romantic fiction of the solitary maker, and that in real institutions responsibility is distributed across many hands. The objection mistakes distribution for dissolution. A film has hundreds of contributors and a structure of authorship nonetheless; a firm has thousands of employees and bears institutional responsibility nonetheless. Authorship can be distributed across the people who set the ends, chose the model, framed the interface, set the standards, designed the verification, and accepted the output, and it remains authorship so long as each of those acts was someone's answerable judgment and the chain can be reconstructed; that the bearer may be an institution rather than an individual is no objection, since groups can be genuine agents to which responsibility attaches (List 2021). What converts distribution into dissolution is not the number of hands but the absence, at one or more of the five points, of any hand at all, the place where the frame was inherited rather than chosen, and no one decided. 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. Refuse both horns. The standard 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. 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. And adoption shows the standard is ordinarily meetable: the physician satisfies all five junctures at working speed because her profession confronted the value choices once, answerably, upstream. 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 So-So Automation and the Direction of AI

The account so far locates abdication in the conduct of individual use: the recruiter who ratifies rather than judges. But the conditions under which she does so are not of her making, and an ethics that stopped at her would mislocate the problem. Whether a system is built and deployed so as to invite authorship or to invite abdication is itself a choice, made upstream, at the level of design and ultimately of economic direction. The most illuminating frame for that level is the task-based account of automation developed by Acemoglu and Restrepo (2018, 2019).

On their analysis, technological change acts on labor through two opposed channels. Automation substitutes machines for labor in tasks previously done by humans, producing a displacement effect that, taken alone, reduces the role and the share of labor. But technology can also create new tasks in which labor has a comparative advantage, a reinstatement effect that restores labor's role in the production process. The history of beneficial mechanization is largely the history of reinstatement keeping pace with displacement: as old tasks were automated, new ones were created in which human work was newly valuable. The pathological case, which Acemoglu and Restrepo (2019) call so-so automation technologies and Acemoglu and Johnson (2023) develop further, is automation that displaces human labor without delivering the large productivity gains or the new complementary tasks that would justify it: technology adopted because it is cheaper than workers rather than because it is markedly better, which lowers labor's share while barely raising output. Self-checkout is their stock example: not obviously more productive than the people it replaces, merely less expensive, and corrosive of the bargaining position of those who remain.

What matters here is not the labor economics as such but its normative shape, and an analogue it suggests at the level of judgment rather than of tasks. Generative AI does not only automate tasks; it can automate the appearance of evaluative work. A system that drafts, ranks, classifies, and recommends can be useful, but it can also remove the human task of forming standards, weighing a case, and deciding what matters, while leaving in place the outward form of a decision having been made. When it does the latter, it effects what I will call a so-so automation of judgment: enough automation to displace the human evaluative task and to speed or cheapen the decision, but not enough preserved authorship to improve, or even to maintain, the quality of responsibility. The recruiter's situation is the microcosm. Her judgment has, in effect, been automated away, not because a machine now judges but because the task that used to require her judgment has been reorganized so that it no longer does, while the records still show a human deciding.

The parallel should be held at the right strength. Acemoglu and Johnson's category is defined by a measurable shortfall, automation that displaces labor without the productivity gain that would justify it, and judgment has no clean analogue of total factor productivity by which a shortfall could be priced. I intend "so-so automation of judgment" as a structural diagnosis rather than an economic measurement: what it displaces is answerability, not output, and what goes uncreated is not GDP but the conditions of responsible action. The nearest existing concept is Danaher and Nyholm's (2021) "achievement gap," the worry that automating a task can deprive a person of the achievement that performing it would have constituted. The answerability gap is its graver relative: what is hollowed is not the achievement but the answerability without which the act has no responsible author at all.

There is more than analogy here. Acemoglu's own forecast turns on the distinction I am pressing. His case for a modest effect, no more than about 0.66% of total factor productivity over ten years

against the multiples the optimists project, rests in its sharpest form on a discount: the productivity evidence so far comes from easy-to-learn tasks, while the tasks carrying the larger promise are hard-to-learn ones, thick with context-dependent judgment and offering no objective outcome measure from which a system could learn to perform them well (Acemoglu 2024). That last condition is my epistemic asymmetry stated in the language of output. A task that gives the system no external yardstick is one it cannot reliably learn, and, for the same reason and at the same point, one whose evaluative frame it cannot author: there is no reality it answers to. Acemoglu withholds the productivity; I withhold the authorship; the two lacks coincide — both arise in tasks with no objective measure — but they are not one lack: his is the absence of a signal the system could learn from, mine the absence of a reality it answers to.

My claim does not, however, need his forecast to hold. Should capability instead far exceed it, that will be because systems found objective yardsticks where he expected none, and a yardstick that lets a system learn a task still supplies no agent who answers for it. The productivity can arrive; the authorship cannot arrive with it. This is why I take Acemoglu's mechanism and not his measure: the number is a claim about output, which may prove low, and the shortfall I name is not a claim about output at all. It survives either outcome.

This reframing carries the same anti-fatalist charge that Acemoglu and Johnson press against the economics of automation. Their central insistence is that the direction of technology is a choice rather than a destiny: the drift toward replacing human work, and concentrating the gains, reflects decisions about what to build and what to build it for, not an inevitability to which we must adapt. Brynjolfsson (2022) presses the point from inside the technology: aiming at human-like AI, systems built to imitate and replace people rather than to extend what they can do, biases the whole enterprise toward substitution, a drift he calls the Turing trap and one easily mistaken for fate. The same is true of judgment. Whether AI is built to displace human evaluative work or to create new and better evaluative tasks, work in which authorship, verification, and care are strengthened rather than hollowed, is a decision made by those who design and procure these systems, and it is the first and most consequential exercise of authorship of ends. To accept that AI must take the judgment-displacing form, because that is simply where the technology is going, is to perform at the level of an industry the same abdication the recruiter performs at her desk: to treat as already settled, by the machine and its momentum, a question that is ours to answer.

The point reaches one level further, to alignment, the governing aim of the field. To align a system is to make its ends good; it is not to make them anyone's, and it cannot be, for the reason §2 gave: being answerable for a set of values is not a property the values can carry on their own, however good they are. Alignment is therefore necessary and not sufficient in just the way control is, securing what the system pursues without securing that a human authored the pursuit. A well-aligned model can leave the answerability gap exactly where a misaligned one does.

7 Three Implications for Design

If the binding problem is abdicated judgment rather than absent control, then design should be evaluated by whether it preserves authorship, not by whether it inserts a human somewhere into a process. Three implications follow directly from the five dimensions of §5, and I state them as design commitments rather than as a comprehensive program; a longer list would only recapitulate the analysis in the idiom of a manifesto.

The first is to separate generation from acceptance. A system may generate a candidate output, but the act by which that output becomes action must remain a distinct and identifiable act, performed by a party answerable for it under standards fixed in advance. In concrete terms, this means resisting designs in which acceptance is the default: the draft sends itself, the ranking is acted on unless someone intervenes, the classification writes itself to the record. Proposals should be legible as proposals until someone accepts them, and the acceptance should be an act rather than the absence of an objection. This is authorship of acceptance, built into the architecture rather than left to the discipline of the user.

The second is to keep the standards outside the model. The criteria by which an output is judged adequate must not be generated by the same system whose output they govern, and must not be silently inherited from the model's own confidence or from the distribution of its training data. They must be specified, and owned, by humans or institutions answerable for them, and held where the model can be measured against them rather than left to certify itself. In the hiring case this is the difference between a system that applies a conception of merit the firm has examined and can defend, and a system that recovers a conception of merit from its data and is then treated as having satisfied it. This is authorship of ends and of standards, made a property of the system rather than a hope about its users.

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: to say who set the goal, who chose the model, who framed the interface, who set the thresholds, who verified the outputs, and who accepted the consequences. A system for which this reconstruction is impossible has not eliminated responsibility; it has hidden it, which is worse. Traceability is what keeps distributed authorship from collapsing into the comfortable anonymity of "the AI decided." The worry long predates generative AI: Nissenbaum (1996) diagnosed the same erosion of accountability in computerized systems decades ago, tracing it to the problem of many hands, the way software faults diffuse responsibility, and the temptation to treat the computer itself as the answerable party. What is new is only the fluency with which today's systems invite that last mistake.

The design side is converging on these commitments independently: Zhu et al. (2026) reach several of them from a human-computer-interaction and governance direction, separating a system's

operative agency in task execution from the human's evaluative agency in verification, steering, and substitution, tying evaluation to criteria external to the system, exploiting the asymmetry between solving a task and checking a solution, and naming the rubber-stamp risk, without the responsibility-theoretic account this paper supplies; the contribution here is the answerability those design moves presuppose, not the layering itself.

A limit must be owned here, and it is the thesis applied to its own remedy: design can scaffold these junctures[3], surfacing the target variable as a contestable choice or forcing a verification step that must be actively cleared, but it cannot manufacture the judgment, since a target variable shown in bright letters can still be waved through and an inserted check clicked past as reflexively as it was added. Authorship is an exercised relation; the most an architecture can do is refuse to let the exercise be invisible. It can never perform it for the user.

A fourth consideration is not a separate principle so much as a caution that cuts across the three. The way a system is presented to its users, the role it is given in the interaction, is ethically consequential because it shapes both what users expect of it and how responsibility is attributed. A system framed as a colleague, an expert, or an oracle invites the deference that hollows authorship and diffuses answerability; a system framed as an instrument keeps the user in the authorial position. I do not claim that any particular framing is always illicit, and warmth, clarity, and accessibility are entirely compatible with honest design. The point is narrower and should be kept free of the moralizing list it usually attracts: anthropomorphic role assignment is not a cosmetic choice, because the role a system is given partly determines whether the human who uses it will author what it does or merely accept it.

8 Objections

Is this not simply Vallor's mirror thesis restated? No. The mirror thesis is a diagnosis: it explains why deference to AI outputs is tempting and why the reflection can be mistaken for authority. I have accepted that diagnosis and built on it. The contribution lies downstream of it, in three claims Vallor's account does not make: that human answerability is independent of the machine's moral status (the decoupling thesis), that the leading framework for preserving responsibility is insufficient for the failure mode generative systems produce (the argument of §4), and that the repair is a specific and design-relevant norm of authorship (§§5–7). Diagnosis and remedy are different achievements, and it is the remedy I am defending.

[3]I have implemented an instance of these commitments in a production website-transformation system, in which disallowed operations (cross-tenant propagation of structural decisions, contamination of one region's content by another, a detector enacting its own repairs) are made structurally inexpressible at the interface level rather than prevented by convention. The architecture is documented in U.S. Provisional Patent Application No. 64/063,345 (filed May 2026), on which I am a named inventor. It scaffolds the junctures; it does not, and on the argument above cannot, supply the judgment exercised within them.

Does the decoupling thesis not fail if AI becomes conscious? This is the objection the thesis is built to withstand, and the answer is no, for the reason given in §2. The arrival of patiency would change what we owe to the system; it would not make the system a fit bearer of the responsibility that the human cannot transfer to it, because the capacity to be wronged and the capacity to answer for an action are distinct capacities, and the second does not come with the first. A stronger version of the objection presses the responsibility gap: a sufficiently autonomous future system, it may be said, would genuinely act on its own and so genuinely break the chain of human responsibility. But the responsibility gap, as §3 argued, is a failure to preserve a human locus, not a transfer of responsibility to the machine; that some systems can be built so that no human answers for them is a fact about how they were built, and an indictment of building them that way, not a discovery that answerability has migrated into silicon.

Does it fail, instead, if AI becomes not a patient but a participant? Suppose a system crosses the threshold §2 left open, becoming a genuine participant in the space of reasons, able to give reasons and to answer for reasons of its own. This is the agency-side twin of the consciousness objection, and §2 was built to concede it rather than to resist it. A participant machine would answer for what it does; it would not thereby author the frame that no one authored, nor take on the account owed for an act that remained the human's, because earning the standing to answer for one's own conduct is not the same as having exercised judgment over the terms of someone else's decision. The recruiter's frame is unauthored whether the system that handed it to her is an instrument, a participant, or something between. Patiency and participation can each run as far as they will; the answerability the human owes does not slide along either, because it never rested on the machine's position on them. Where present and future systems fall on those gradients, and what becomes of answerability when participants of a genuinely alien kind stand in the chain, is a question this paper leaves open and pursues elsewhere; the floor it sets is that the answer, wherever it lands, leaves who authored the frame exactly where it found it.

Is the problem not responsibility abundance rather than a gap? Kiener (2025) argues, against the gap tradition, that AI-mediated harms usually leave us with too many attributable parties rather than too few, since developers, deployers, configurers, users, and managers can all be connected to the outcome, so that the real difficulty is adjudicating the abundance. I do not dispute the abundance; as §4 conceded, the hiring case has attributable parties in plenty. But abundance is a fact about attribution, about how many parties can be connected to the outcome, and is orthogonal to the answerability question pressed here. A decision can have a crowd of attributable parties and still be one in which no one answerably authored the evaluative frame, because each took it ready-made from the next: the deployer from the vendor's default, the user from the deployer's configuration, the vendor from the training distribution. The answerability gap is therefore not a rival to Kiener's abundance but its complement. Abundance describes who can be held to account;

the answerability gap describes the judgment that none of them performed. Together they explain why blame so often settles on the connected human who authored least, the front-line approver who becomes, in Elish's (2019) phrase, a moral crumple zone, the zone here opened by a deficit of judgment rather than of control, while the appearance of deliberation, supplied by the sheer number of hands, conceals that no hand deliberated.

Does privileging human judgment not ignore that human judgment is itself biased? It does not, and the objection, taken seriously, strengthens the account. The bias in the hiring system did not originate in the machine; it was inherited from a human history encoded in the data. The claim of this paper is not that human judgment is reliable or unbiased, since it is neither, but that responsibility requires an answerable party who has actually exercised judgment, and that the right response to bias is to author more carefully, not to abdicate. A system that launders inherited bias through a fluent ranking while removing the human who might have interrogated it has not corrected for human fallibility; it has automated it and erased the party who could be held to account for it. Authorship is the demand that someone answer for confronting the bias, which is precisely what abdication evades.

Is this not, in the end, a counsel against using AI? It is the opposite. The argument grants that these systems can be useful, locates the danger not in automation but in a particular direction of automation, and holds that the corrective is to aim AI at augmenting human judgment rather than displacing it. Authorship constrains where automation may run without remainder; it does not oppose automation. A system designed so that its users author the ends, standards, verification, acceptance, and final form of what it produces is not a diminished system but a better one, and building such systems is the practical upshot of everything argued here.

9 Conclusion

Whether these systems are, or might become, minds we could wrong is a real question, and I have not settled it; I have argued only that it is the wrong question to put at the center of the ethics of using them, because its answer, whatever it proves to be, leaves human answerability untouched. Patiency is uncertain and may change; answerability is immediate and does not. The work on responsibility gaps and meaningful human control teaches us to demand a human connected to the system closely enough to be held to account, and that demand is sound; but for the fluent generative and decision-support systems now woven into institutional life, connection is not the binding constraint. We can usually name the human who signed off; what we cannot assume is that she authored what she signed. The mirror is why this is tempting; the gap is why it is dangerous; authorship is the norm that closes it, and the direction we give the technology is where it is honored or abandoned at scale.

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.

References

Acemoglu, D. (2024). *The simple macroeconomics of AI* (NBER Working Paper No. 32487). National Bureau of Economic Research. https://doi.org/10.3386/w32487

Acemoglu, D., & Johnson, S. (2023). *Power and progress: Our thousand-year struggle over technology and prosperity.* PublicAffairs.

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review, 108*(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives, 33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

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

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

Brynjolfsson, E. (2022). The Turing trap: The promise and peril of human-like artificial intelligence. *Daedalus, 151*(2), 272–287. https://doi.org/10.1162/daed_a_01915

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

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

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

Danaher, J., & Nyholm, S. (2021). Automation, work and the achievement gap. *AI and Ethics, 1*(3), 227–237. https://doi.org/10.1007/s43681-020-00028-x

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

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility.* Cambridge University Press.

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

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

Lerchner, A. (2026). *The abstraction fallacy: Why AI can simulate but not instantiate consciousness* [Preprint]. PhilArchive. https://philarchive.org/rec/LERTAF

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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI welfare seriously.* arXiv. https://doi.org/10.48550/arXiv.2411.00986

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

Moran, R. (2005). Getting told and being believed. *Philosophers' Imprint, 5*(5), 1–29. https://hdl.handle.net/2027/spo.3521354.0005.005

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

Nawar, T. (2024). Generative artificial intelligence and authorship gaps. *American Philosophical Quarterly, 61*(4), 355–367. https://doi.org/10.5406/21521123.61.4.05

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics, 2*(1), 25–42. https://doi.org/10.1007/BF02639315

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

Searle, J. R. (1992). *The rediscovery of the mind.* MIT Press.

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

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

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/

Thaler v. Perlmutter, 130 F.4th 1039 (D.C. Cir. 2025), cert. denied, No. 25-449 (U.S. Mar. 2, 2026).

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

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

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

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet, 11*(1), 104–122. https://doi.org/10.1002/poi3.198

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics, 24*(2), 227–248. https://doi.org/10.5840/philtopics199624222

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* W. H. Freeman.

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

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-poisonedrag