The Authority Gap

Why Intelligence Alone Cannot Govern Action

Bradley D. Wilson

20 October 2026

Abstract

An artificial intelligence (AI) system can produce a correct decision without being entitled to issue it. More subtly, a permission, an approved document, and a delivery record can each be valid while referring to different versions of an institutional act. Adequate governance requires an institution to connect its mandate to the materially relevant action and its consequences. Through conceptual analysis and contrasting hypothetical cases, I develop this relational account of the authority gap. Five diagnostic dimensions organize the analysis: interpretation, authorization, execution, evidence, and recovery. The account distinguishes an unauthorized act from an authorized act whose status cannot be established, and both from an act that faithfully implements an unjust rule. The analysis of recovery shows why permission for an original action does not automatically settle permission for another attempt when the outcome is unknown. Continuity of authority grounds these requirements in obligations to affected people that survive changes in how institutional work is performed. It supports standing mandates and justified action under uncertainty while requiring assurance proportionate to those obligations. Building on responsibility gaps, reviewable automation, authenticated delegation, and bounded autonomy, I propose governed execution as a field of inquiry into these relationships. Its central task is to explain how institutions can preserve the authority for their actions through performance, account for uncertain effects, and remain answerable to the people affected.

Keywords: AI governance; authority; delegation; institutional action; human oversight; accountability

1 Introduction

Intelligence does not confer authority. Consider a benefits office that uses an AI assistant to prepare correspondence. A caseworker asks it to finish a notice concerning a claimant’s eligibility. The assistant identifies the relevant file, produces a clear explanation, and sends the notice through the office’s messaging service. The explanation may be factually accurate. The difficulty is that the caseworker intended to review a draft, the notice triggers a deadline, and the recipient has reason to treat it as an official decision. Nothing in the quality of the writing settles whether the system was entitled to send it.

The problem becomes harder when every component works. An identity service can authenticate the caseworker, a review system can record approval of one draft, and a messaging service can faithfully transmit a different version. Each record is accurate. Together, they fail to establish that the notice sent was the notice authorized. If the delivery response then disappears, the office faces a further question: does its authority to issue the notice permit another attempt when the first may already have taken effect?

These are the two problems at the center of this paper. Adequate governance requires a defensible connection between the mandate and the materially relevant act. That connection must also govern

recovery when the institution cannot establish what happened. The resulting obligations concern the relationships among components, and cannot be discharged simply by certifying each component separately.

Institutional work depends on distinctions between advice and decision, preparation and commitment, and the possession of information and the standing to act on it. AI systems can obscure those distinctions because the same interface may interpret a request, formulate a plan, select a recipient, and initiate an operation. ReAct, Toolformer, and SWE-agent illustrate different ways of connecting language-model capabilities to tools and environments (Yao et al., 2023; Schick et al., 2023; Yang et al., 2024). Their technical achievements make the institutional question more pressing: which of those operations has anyone authorized?

There is no clean boundary between harmless text and consequential action. A statement can give notice, make a promise, disclose confidential information, or induce someone to change their conduct. Conversely, a tool call may merely retrieve a public document. The relevant question is what the operation does in its setting. In this paper, an institutional action is an operation or communication that exercises, represents, or materially affects an institution's powers, commitments, records, or relationships with people. Its consequences may depend on formal rules, reasonable reliance, or both.

The problem predates contemporary AI. Bainbridge (1983) examined the difficulties created when people are left to supervise processes whose routine operation has been automated. Language-model agents extend this problem by interpreting broad instructions and constructing much of the route from instruction to consequence. That route can contain consequential choices that nobody settled when access was granted.

I use authority gap to describe a failure to establish or maintain the relationship between an action and the authority claimed for it. This includes cases in which authority is absent, cases in which a valid delegation is exceeded, and cases in which the available evidence cannot establish whether an action remained within its mandate. These are different failures. An action does not become unauthorized merely because its records were lost, although the loss may prevent the institution from justifying what it did.

Institutions owe people more than competent performance. Where an institution owes an authorized determination, a fair opportunity to challenge it, or a remedy for an error, delegation must preserve the conditions for meeting those obligations. The institution must provide controls and evidence adequate to connect its mandate to what it does. I call this requirement continuity of authority. Section 8 defends it by examining why a change in the means of institutional action cannot, on its own, release the institution from its duties.

Continuity allows the means of work to change. A benefits office may replace case-by-case dispatch approval with a justified standing mandate. It may not treat that change alone as a reason to withdraw a claimant's opportunity to challenge an erroneous notice. The content of those obligations comes from the institution's relationship with the claimant; the technology must be assessed against them.

Governed execution names the practice and proposed field of inquiry organized around these requirements. Its subject is the exercise of authority through actual human and machine operations, including what happens when those operations produce disputed or uncertain effects. I defend a broad remit because justification and recovery constrain the same institutional act. A technically reliable

dispatch can still lack authority, and a technically sensible retry can exceed the authority that covered the first attempt.

The method is conceptual and normative. The benefits office and its variations are constructed cases that separate questions commonly collapsed into a single judgment of system reliability. I use selected philosophical, technical, and empirical literature to examine those distinctions and their consequences. Sections 2 to 4 establish the concepts and their relationship to prior work. Sections 5 and 7 develop the central analyses of authorization across components and recovery under uncertainty; sections 6 and 8 examine human control and the ethical grounds of continuity. Sections 9 to 11 draw out the design principles, research agenda, and limits of the account.

2 Intelligence and authority

Intelligence, causal power, and authority answer different questions. Intelligence concerns capacities such as reasoning, interpretation, and adaptation. Causal power concerns the ability to bring about an effect. Authority concerns a standing to act, decide, or direct within a relationship governed by reasons and rules. These capacities can coincide in one actor, but none should be inferred automatically from the others. A highly competent adviser may lack decision-making power. An administrator may possess a credential that permits an operation the administrator has no right to perform.

Authority involves several distinct questions. A technical permission is a system's rule allowing an operation. Institutional authority is a recognized competence or delegation within an organization or social practice. Legal validity concerns the effect that law gives an act. Moral legitimacy concerns whether the exercise of power is justified. Evidence concerns what can be established about any of these matters. An action can satisfy one condition and fail another. A policy engine may permit a discriminatory decision; an unauthorized communication may nevertheless expose its institution to liability; a valid decision may be poorly documented.

Here, institutional authority means a recognized standing to perform a class of acts under specified conditions. I use legitimate authority when the moral justification of that standing is at issue. Recognition within an institution does not settle whether its rules deserve compliance. A system may establish that it obeyed those rules while leaving their justification entirely unresolved. Governed execution must therefore permit scrutiny of the mandate as well as scrutiny of its performance.

Searle's account of institutional facts helps explain why a familiar form is insufficient to establish an act's status (Searle, 1995). A document becomes an official determination through the rules and relationships in which it is produced, rather than through its resemblance to other determinations. Applied to the benefits example, the relevant questions include which office may decide the matter and what procedures give the notice its force. A model's ability to reproduce the genre does not answer them.

Austin's treatment of speech acts makes the same example more demanding (Austin, 1962). Language can participate in doing something, and the circumstances of an utterance help determine what has been done. An institution therefore cannot assess an assistant solely by asking whether its sentences are true. It must also ask what those sentences count as when delivered through an official channel. A notice that accurately describes a rule may still be premature, issued by the wrong office, or addressed to someone whose case has not been decided.

Weber's discussion of legitimate domination concerns the beliefs through which a relation of rule is sustained (Weber, 1978, Part I, Chapter III). Raz asks a different question: under what conditions does accepting an authority's directives help a person conform to reasons that already apply to that person (Raz, 1986, Chapter 3)? The difference matters here. An official-looking communication may be accepted because of the office it appears to represent, while the institution's entitlement to issue it remains disputed.

Raz also supplies an objection to an easy dismissal of machine authority. His normal justification thesis compares following an authority with attempting to follow the applicable reasons directly, not simply with following a less accurate human adviser. Applied to AI, it invites an argument that accepting a system's directives could sometimes improve conformity with those reasons. I accept that possibility. It supports a role whose scope must be justified by the relevant reasons, while leaving institutional recognition and moral justification open to separate assessment. The question is what standing is warranted, rather than whether superior performance entitles a system to act without limits.

Natural language is not inherently incapable of conferring authority. A competent principal may give a valid instruction orally or in writing. The difficulty for an automated system is establishing who gave the instruction, what that person could authorize, and how its meaning applies to the proposed act. A structured authorization can make these questions easier to inspect, but structure is neither a universal legal requirement nor a guarantee of validity. The institution remains responsible for the connection between the representation and the authority it represents.

3 Five dimensions of the authority gap

I distinguish five dimensions of the authority gap: interpretation, authorization, execution, evidence, and recovery. They follow the questions an institution must answer about an act: what was proposed, what permitted it, what occurred, what can be established, and what may happen next. The dimensions overlap because the answers constrain one another. Recovery, in particular, can reopen every earlier question. Table 1 summarizes them.

The interpretation dimension begins with the translation from a request to a consequential operation. "Finish the notice" leaves open whether the assistant should draft, save, submit for approval, or send. It may also leave unresolved which recipient, version, or channel is intended. Human workplaces manage some of this ambiguity through shared practice and the opportunity to ask questions. An automated system needs a defensible way to use that context. Where different interpretations would change the institutional consequence, silently selecting the most action-oriented interpretation is difficult to justify.

A workable proposal should identify the features that matter to the decision. In the benefits example, these include the case, the status of the underlying determination, the recipient, the notice's material contents, and the consequences of issue. The system need not turn every request into a lengthy form. A standing instruction can settle routine choices. The requirement is that consequential ambiguity be resolved before the institution acts on the interpretation, or be handled under an explicit authority to exercise discretion.

The authorization dimension asks whether the proposed act falls within a valid grant of power. Identity and authorization are related but different. Knowing which account requested an operation does not

establish what its holder may decide. Nor does the ability to use a messaging service establish authority to make every representation that could be sent through it. Attribute-based access control can express rich conditions involving the subject, object, operation, and environment (Hu et al., 2014). The question is whether the conditions actually represented in a particular system cover the institutional distinctions at stake.

Authorization also depends on relationships among acts. A manager may authorize notices only after a separate eligibility decision. A spending mandate may have a cumulative ceiling. Two individually permissible operations may jointly exceed that ceiling, and two different agents may act on the same remaining balance. An adequate account therefore needs the logical action, its relevant history, and any shared limits. Examining isolated API calls will miss some of the most important constraints.

The execution dimension concerns what happens after permission has been established. A reviewer may approve one recipient while a later component substitutes another. A system may generate a new attachment after the review. A subagent may reach an alternative service with different effects. In these cases, the authorization can be valid and the execution still exceed it. Preserving authority requires a relationship between the reviewed terms and the performed operation that survives the relevant transformations.

This does not mean that every byte must remain unchanged. Formatting, transport encoding, or a choice among expressly permitted routes may leave the authorized act intact. Materiality depends on the mandate. Changing a payment recipient is normally different from changing a transport header; changing punctuation may matter greatly in a contractual clause. The appropriate rule is to bind the operation to its material terms and its permitted range of variation. If that range is exceeded, the system needs renewed authorization or an authorized procedure for resolving the change.

The evidence dimension concerns the support for claims about the action. What matters about a log is which event it records, how reliably it records that event, and what inference the institution draws from it. A service’s acknowledgment of receipt may establish that a request arrived without establishing that a notice was delivered or that a recipient read it. Evidence can justify an institutional conclusion without providing certainty; its sufficiency depends on the claim and the consequences of error.

The recovery dimension appears when the account of the result is incomplete or when correction becomes necessary. A communication failure can occur after a remote operation has taken effect. This is a longstanding distributed-systems problem; the RPC specification, for example, explains why the absence of a reply does not determine whether a remote procedure executed (Thurlow, 2009). Its institutional significance depends on the act. Repeating a delivery request may be harmless under a reliable deduplication arrangement, or it may issue a second notice with a new deadline. Recovery must be authorized in light of those possibilities.

The dimensions are useful only if they improve diagnosis. Suppose the notice was validly authorized and correctly sent, but the local record disappeared. The immediate defect is evidential, although it may complicate recovery. Suppose instead that an unauthorized notice was sent and impeccably recorded. Better evidence has exposed the authorization failure rather than repaired it. Treating both cases as a generic loss of trust would conceal the different remedies they require.

Public cases give these distinctions practical stakes. Lifshitz and Hung’s (2024) account of Moffatt v. Air Canada describes liability for negligent misrepresentation through a chatbot. The case illustrates

the consequences of institutionally presented communication; it does not establish a general rule conferring authority on autonomous agents. The Robodebt Royal Commission documents failures in the legality and administration of an automated public scheme (Royal Commission into the Robodebt Scheme, 2023). Both direct attention to the institution's conduct, including the rules it applies and the representations on which people rely. The constructed cases below isolate particular relationships within that wider problem.

Table 1 Five questions about an institutional action

Dimension	Question	Illustrative failure
Interpretation	What act is actually proposed?	An instruction to prepare a notice is interpreted as permission to issue it.
Authorization	Who may permit this act under these conditions?	A valid account is used outside the holder's delegated role.
Execution	Does performance remain within that permission?	The recipient or material content changes after review.
Evidence	What does the record establish?	A request receipt is presented as confirmation of the intended effect.
Recovery	What may be done after failure or uncertainty?	A second attempt creates an additional effect outside the mandate.

4 What existing approaches already provide

Governed execution draws on established work in security, accountability, and the philosophy of responsibility. Clark and Wilson (1987) made controlled transformations and separation of duties central to computer integrity. Nissenbaum (1996) examined obstacles to accountability in computerized systems, including dispersed responsibility and the tendency to blame the computer. These foundations make it possible to ask how authority survives the combination of operations across a workflow.

The responsibility-gap literature is especially important. Matthias (2004) argues that learning systems can undermine the conditions for attributing responsibility to their makers or operators. Santoni de Sio and Mecacci (2021) distinguish culpability, moral accountability, public accountability, and active responsibility, and connect these to meaningful human control. Their analysis reaches organizational and social problems as well as technical unpredictability. It provides a broader account of responsibility within which the present question of authority belongs.

Authority identifies a different question within that wider problem. Suppose an office knowingly sends a notice it was not entitled to issue and accepts full responsibility for doing so. No uncertainty about the responsible institution closes the defect in authorization. Conversely, a notice may be authorized and correctly sent even though a later investigation cannot fairly attribute blame to a particular employee for an unforeseen downstream harm. These cases separate entitlement to act from culpability for consequences. They do not place authority outside accountability: public and active responsibility already require attention to how powers are exercised. This paper specifies the connection between a

mandate and its performance, including cases in which that connection fails despite identifiable responsibility.

Risk-management frameworks address an organization's processes for identifying and managing risks. The AI Risk Management Framework of the National Institute of Standards and Technology (NIST) and ISO/IEC 42001 provide important institutional foundations (NIST, 2023; ISO/IEC, 2023). A deployment must translate those commitments into effective conditions on actual operations. A risk register can identify unauthorized disclosure; an execution design must also determine which disclosure is proposed, whose authority applies, and how it will be constrained.

Alignment research tackles another part of the problem. Cooperative inverse reinforcement learning studies assistance under uncertainty about human preferences, while corrigibility research examines systems that remain amenable to correction (Hadfield-Menell et al., 2016; Soares et al., 2015). These are relevant to institutional use. A system that resists correction or pursues an incorrectly specified objective presents obvious difficulties for delegation. Yet willingness to follow a person does not settle whether that person may direct the act, whether someone else's rights constrain it, or whether an earlier instruction remains applicable.

There is a stronger alignment objection: perhaps a sufficiently complete account of the objective would already include all those restrictions. In that case, the conceptual opposition largely disappears. The remaining demand is assurance. An institution still needs grounds for believing that the system identified the applicable restrictions and remained within them on this occasion. Governed execution describes that problem of accountable performance. It does not show that every suitable assurance must be provided by a component external to the model, although independent controls can reduce reliance on the same system that proposes the action.

Authentication and authorization protocols offer particularly concrete foundations. OAuth Rich Authorization Requests can carry structured details about a requested authorization, including payment-related parameters (Lodderstedt et al., 2023). Access-control infrastructure can therefore express much more than a coarse permission to use a service. The institutional question is whether the authorization record accurately expresses the mandate, constrains the eventual operation, and can be connected to a reliable account of its outcome.

Recent agent research directly addresses delegation. South et al. (2025) argue for authenticated delegation using existing identity and authorization infrastructure. Jørgensen and Ma (2026) integrate bounded autonomy with delegated authority, controlled execution, provenance, and recovery. Their framework substantially overlaps the operational concerns examined here, including the relationships among controls. I build on that work by developing an argument about what those relationships must establish. Valid permissions and successful component operations can accompany an act that exceeds the mandate. An unchanged initial mandate can also require different recovery decisions when the possible effects of another attempt differ. Sections 5 and 7 examine those contrasts, and section 8 explains why they matter to people affected by institutional power.

Legal and accountability research also changes the comparison. Kolt (2026) brings agency theory and law to the governance of AI agents, including the problems of discretion and monitoring. Cobbe and Singh (2020) argue for the reviewability of automated decision-making as a process, rather than treating the model's explanation as the whole account. These approaches already recognize that

responsibility extends beyond a single output. An action-centered practice should build on them and test where their institutional requirements need additional operational expression.

Security work on prompt injection supplies a different reason for binding authority carefully. Greshake et al. (2023) investigate how instructions embedded in external material can affect LLM-integrated applications; AgentDojo provides an environment for evaluating attacks and defenses involving tool-using agents (Debenedetti et al., 2024). Such work shows why information encountered during a task cannot simply acquire the status of an instruction from the principal. Defeating an adversary leaves a further question: did the user have standing to give the instruction in the first place?

OWASP's Agent Control Standard is another recent effort to support portable controls enforced during agent operation (OWASP Gen AI Security Project, 2026). The question for this account is how such arrangements connect a defensible mandate to the operation, the evidence, and the available remedy. Different systems can satisfy that requirement through different combinations of controls.

This paper develops a relational account of authorized institutional action and its recovery. It identifies what must connect a mandate, an operation, and the evidence offered for that operation, then shows how uncertainty about the outcome changes the justification required for further action. Continuity of authority supplies the ethical basis: an institution's duties to affected people constrain the relationships it must preserve. This joins the argument about why institutions owe an account of their actions to the practical question of what that account must establish.

Table 2 locates this contribution within the existing work. The comparison concerns the argument developed here; it does not presume that adjacent frameworks cannot accommodate it. The test is whether this account clarifies the disputed status of an act or changes the justification for what happens next.

Table 2 Relationship to established and adjacent work

Approach	Contribution already available	Task pursued in this paper
Responsibility gaps	Matthias (2004) and Santoni de Sio and Mecacci (2021) examine responsibility, accountability, and control.	Distinguish entitlement to act from attribution of responsibility and specify their relationship.
Integrity and authorization	Clark and Wilson (1987) address controlled transformations and separation of duties. Hu et al. (2014) and RFC 9396 (Lodderstedt et al., 2023) support contextual and detailed permissions.	Explain why valid local controls are insufficient unless they connect the mandate to the material act.
Authenticated delegation	South et al. (2025) connect agent delegation to authentication, scoped permissions, and accountability.	Examine how a mandate applies to the act performed and to further attempts under uncertainty.
Enterprise bounded autonomy	Jørgensen and Ma (2026) integrate authority, execution, evidence, and exception handling.	Develop the relational argument through contrasting cases of valid controls, changed acts, and uncertain effects.

Approach	Contribution already available	Task pursued in this paper
Reviewable decisions	Cobbe and Singh (2020) make the wider decision process an object of review.	Relate review to the particular mandate, material operation, uncertain outcome, and remedy.
Agency law and theory	Kolt (2026) examines discretion, monitoring, and institutional governance of AI agents.	Develop a general normative account while leaving legal conclusions to the applicable jurisdiction.

5 Delegation and the continuity of authority

Delegation permits one actor or system to act within a mandate established by another. For institutional purposes, a mandate should identify its source, the relevant class of acts, the conditions under which those acts may occur, and the discretion left to the delegate. Some mandates are narrow and short-lived. Others establish a continuing role. Neither form should be confused with an unrestricted transfer of the principal's position.

The authority to delegate must itself be established. A caseworker might be entitled to draft a notice but unable to authorize its issue. A manager might be able to authorize routine notices but unable to waive a statutory condition. A user who controls an agent's configuration does not necessarily hold every institutional power the configured tools can exercise. The point is especially important when a personal assistant operates across services with different owners, purposes, and rules. The user's desire to complete a task is relevant, but it is not the only source of obligations.

A delegation should also distinguish discretion from uncertainty about the instruction. Giving an assistant a choice among approved delivery channels is a grant of discretion. Failing to state whether it may disclose a medical detail is an unresolved boundary. The two can look similar to a planning system, since both require a choice, but they have different justificatory structures. In the first case, the principal has authorized a range. In the second, the system may be inferring permission from silence.

An institution can make this distinction operational by specifying what is fixed, what may vary, and what must be referred back. For the benefits notice, the recipient and case determination might be fixed; routine formatting might be discretionary; a conflict between the file and the proposed explanation might require review. The representation should be intelligible to the people responsible for the decision as well as enforceable by the system. If those audiences understand it differently, a formally valid token may conceal a disagreement about the act.

Delegation through several agents compounds the problem. An orchestrator may assign research to one component, drafting to another, and transmission to a third. A downstream component should not obtain a broader mandate merely because an upstream component finds broader access useful. The appropriate default is attenuation: inherited authority remains within the limits of its source. If additional authority is genuinely needed, an actor entitled to grant it must do so. That new grant should be distinguishable from the original delegation rather than silently attributed to it.

Limits may concern combinations of operations as well as individual operations. An agent authorized to spend within a budget cannot evade that budget by dividing work among subagents. Similarly, a limit on disclosure applies to the information released across a sequence, not merely to each message

viewed separately. A system that checks each component in isolation may preserve local permissions while defeating the purpose of the mandate. Continuity therefore includes relevant obligations carried across the chain.

Revocation, expiry, consumption, and invalidation require separate treatment. Revocation withdraws a grant; expiry ends it under a time condition; consumption may exhaust a one-use permission; invalidation may result from a change in the conditions that made it applicable. These events need not have the same legal or operational effect. A validly completed action may continue to bind the institution after a mandate expires. A system must identify the decision point at which a change to authority affects an operation that is already under way.

Checking authority immediately before consequence is harder than the phrase suggests. Remote systems do not always provide a single, observable instant at which authorization and effect can be made atomic. The interval between a local check and a remote commitment may matter. A responsible design should state where the binding check occurs, what changes can still intervene, and what happens to requests in flight. Calling a gateway non-bypassable does not establish these properties; the claim depends on the actual routes by which the relevant state can change.

The proposed lifecycle keeps these questions connected. Proposal identifies the act. Governance determines the applicable rules and the reasons they express. Authorization establishes that the proposed operation falls within an available mandate. Execution applies that mandate to performance. Evidence supports an account of the result. Recovery addresses correction, reconciliation, or further action. These are functions, not necessarily six separate services or a sequence traversed only once.

Table 3 Responsibilities across the action lifecycle

Function	Institutional responsibility
Proposal	Describe the intended act and resolve consequential ambiguity.
Governance	Identify applicable rules, rights, and obligations, including conflicts requiring judgment.
Authorization	Establish a valid mandate for the act and its permitted discretion.
Execution	Keep performance within the mandate through the relevant paths and transformations.
Evidence	Preserve evidence adequate to support claims about authorization, attempts, and effects.
Recovery	Determine what correction or further action is justified given the current evidence.

The lifecycle may begin with a standing authorization and move through many routine acts. It may also return to governance when a case reveals an exception. A human can perform several of its functions. The criterion is whether the relationships hold: completing an approval step and an execution step does not establish that the approved operation was the one performed.

Consider the office introduced in section 1. Its identity service correctly authenticates the caseworker, its review system accurately records approval of a draft, and its messaging service reliably sends the file it receives. The caseworker has authority to approve issue of the notice, but the assistant has no discretion to alter its material content after approval. The assistant’s local specification requires a valid recipient and a notice in the required format; it does not require comparison with the approved draft. Between review and dispatch, the assistant changes a paragraph that affects the claimant’s deadline.

Every component meets its local specification. The institution nevertheless issues a notice that the approval did not cover.

Improving the accuracy of those three records would not repair the defect. They already report their respective events correctly. What the office lacks is a constraint connecting approval to the content sent, and evidence that the constraint held. A receipt for the altered notice cannot make the earlier approval apply to it. The case establishes a requirement on the relationship among controls: their adequacy must be assessed against the same act in the respects made material by the mandate.

The converse matters too. An automated dispatch can satisfy a valid standing mandate without recording a separate human approval for each notice. If the arrangement establishes membership in the authorized class, preserves the material terms, and supplies evidence adequate to the office's obligations, the absence of an additional approval record is not a gap. The first office has an approval record but sends an unauthorized notice; the second can act with authority without that record. Continuity therefore tests the relationships required by the mandate rather than counting approval steps.

Apparent authority raises a further issue: an institution's public presentation can affect what others reasonably believe about a system's role. Agency doctrine distinguishes actual and apparent authority in ways that depend on the relevant legal system and facts (American Law Institute, 2006). A branded assistant's statement may create reliance or liability without giving the assistant every power it appears to possess. Institutions should align the role they present with the mandate they enforce, and give affected people a workable route to clarification and remedy.

6 Human oversight and nondelegable responsibilities

Placing a human somewhere in a process does not, by itself, establish meaningful control. The person may lack authority, information, time, or a practical means of intervention. The opposite assumption is equally mistaken: meaningful human control need not require a person to approve each operation. Santoni de Sio and van den Hoven (2018) connect control to responsiveness to human reasons and the ability to trace responsibility to people with an appropriate understanding of their role. Davidovic (2023) further emphasizes that the purpose for which control is required affects what counts as adequate control.

The empirical literature gives reason to treat supervision as a design problem. Parasuraman and Riley (1997) distinguished several ways automation can be used poorly, including overreliance and underuse. In a simulated flight task, Skitka et al. (1999) found that participants using an imperfect automated aid made both omission and commission errors. In that setting, access to contradictory information did not prevent those errors. An institution must ask whether its review arrangement actually enables people to identify and act on a discrepancy.

More recent evidence complicates the assumption that adding a person necessarily improves performance. Vaccaro et al. (2024) reviewed 106 experimental studies and found that human–AI combinations, on average, performed worse than the better of the human or AI conditions alone, with substantial differences among tasks. That finding concerns performance under the studied arrangements, not the legitimacy of delegating an institutional decision. It nevertheless supports a

practical caution: an oversight requirement needs an account of what the human contributes and an evaluation of whether the arrangement enables that contribution.

Romeo and Conti's review (2026, first published online in 2025) examines automation bias in human–AI collaboration. Laux and Ruschemeier (2025) analyze the difficulties of translating concern about such bias into legal duties under the AI Act. Taken together with the older literature, these works support an approach attentive to interface design, workload, and organizational setting. A warning telling reviewers to remain vigilant is unlikely to settle the question of whether they can exercise independent judgment.

Where an act requires individual review, I propose five conditions for assessing it. The reviewer must hold the relevant authority. The information presented must be adequate for the judgment required. Review must occur while intervention can still make the intended difference. The reviewer must have usable options to reject, revise, or escalate. Finally, the resulting decision must constrain what the system actually performs. A failure in any of these conditions gives the institution a specific defect to address.

Adequate information is not the same as maximum information. A reviewer facing hundreds of unprioritized fields may be less able to identify a material change than one shown a concise, accurate comparison. The interface should make clear the intended effect, the reasons for acting, the material terms, and any departure from a prior authorization. Where the system offers an explanation, the explanation should not substitute for the underlying evidence needed to assess the decision. The reviewer must be able to inspect relevant details without being forced to reconstruct the entire technical process.

The organizational conditions matter just as much. A nominal veto has little practical force if using it is routinely penalized, if the workload makes examination impossible, or if responsibility is assigned to someone who cannot change the system. Elish (2019) describes the misallocation of responsibility to operators with limited control as moral crumple zones. Responsibility should follow control over the relevant decisions, resources, and design choices. An institution cannot establish meaningful oversight by assigning blame to the last person near the machine.

Standing authorization can provide meaningful control over routine activity when the permitted class is sufficiently well specified and the arrangement remains open to review. The benefits office might approve an established notice template, a defined class of completed determinations, and rules for selecting recipients. Individual dispatch need not then await a fresh human decision if all conditions are met. Cases outside the standing mandate should follow an appropriate exception process. That arrangement can preserve institutional judgment without turning every permitted act into an approval request.

Its adequacy depends on how the institution establishes that a case belongs to the permitted class, detects changes that defeat the mandate, and enables an authorized person to revise or withdraw it. Repeated manual approval can add cost without repairing deficiencies in those conditions.

Some responsibilities nonetheless resist delegation. Three grounds should be distinguished. A legal rule may reserve a decision or act to a specified office or professional. A moral argument may require personal judgment, participation, or accountability even where the law permits automation. A prudential restriction may reflect the current system's inability to provide adequate assurance. The

third ground may change as evidence improves; the first requires interpretation of the applicable law; the second requires an argument about the people and interests at stake.

The retention of responsibility does not entail that every component of performance must remain manual. A professional may retain a duty while using automated preparation or transmission. Conversely, a system may execute an operation flawlessly while participating in a decision that should never have been delegated to it. For this reason, nondelegability should be specified at the level of the relevant judgment, commitment, or act. Broad lists of sectors cannot substitute for that analysis.

The law also supplies clear counterexamples to a universal requirement of contemporaneous human approval. In the United States, 15 U.S.C. § 7001(h) addresses the validity of records and contracts involving electronic agents when their actions are legally attributable to the person to be bound (United States, 2000). This does not authorize every automated transaction, but it shows why the mere absence of a human decision at the moment of execution cannot establish invalidity. The required authority may have been conferred earlier.

For rights-affecting and difficult-to-remedy decisions, a stronger case can often be made for participation, explanation, and review. Those safeguards must be fitted to the relevant context. A payment may be recoverable while the deprivation it caused was not; a software deployment may be rolled back while a disclosure remains permanent. The relevant question is the reversibility of the consequences for people, not simply whether a system offers an undo function. An institution should explain both the limits it imposes on automation and the reasons those limits are adequate.

7 Evidence and responsible recovery

An institution’s account of an action should distinguish what was intended, attempted, observed, and established. These categories are easily collapsed by an interface that reports a single status such as “complete.” A request can be accepted before processing finishes; a local failure can occur after a remote commitment; a sent message can remain undelivered. If the institution reports a stronger conclusion than its evidence supports, it creates another governance problem even when the underlying operation was authorized.

Useful evidence depends on the claim under examination. To establish that a proposal received approval, a record should connect the reviewer, the material terms, and the decision. To establish that an operation occurred, the relevant evidence may come from a provider, an authoritative register, or an observed change in state. To establish that the approved act and the performed act correspond, the institution needs a reliable connection between those records. No single receipt necessarily establishes all three propositions.

The reliability of that connection depends on assumptions that should be visible. Records can be incomplete, clocks can disagree, identifiers can be reused, and observers can fail. A signature may establish that a particular key signed a statement without establishing that the statement is true. An append-only record can preserve an error faithfully. The institution’s claims must therefore match what its recording process can establish and, where the consequences warrant it, draw on independent observations or reconciliation.

Evidence of non-action is especially demanding. The absence of an entry in one log does not establish that no effect occurred through another path. A defensible non-action claim therefore needs a

boundary: which operation, through which possible routes, during which interval, under which assumptions about observation and enforcement. A system may establish that a proposed notice never crossed a controlled dispatch boundary. It should not turn that bounded finding into a claim that nobody, by any means, communicated with the claimant.

For the benefits office, this distinction changes the recovery question. Suppose the office submits an authorized notice and loses the response. It cannot establish whether dispatch occurred. As the RPC specification makes clear, failure to receive a reply does not establish that the remote operation was not executed (Thurlow, 2009, section 5). The office still has a valid mandate to issue one notice. What it lacks is a basis for deciding whether another transmission would complete that act or produce an additional effect.

Consider two services. One reliably treats a repeated request with the same identifier and material content as the same dispatch, within its documented conditions. The other treats each accepted request as a fresh notice, potentially starting another deadline. The original mandate can support a further attempt with the first service if it covers that behavior and the conditions hold. It does not, merely by authorizing one notice, authorize the possible additional effect with the second. The technical difference changes what the institution must justify before acting again.

An unknown outcome need not bring all action to a halt. It may be safe to repeat an idempotent operation under an authorization that covers that behavior. It may be necessary to issue a clarification, contact the affected person, or take a protective action with bounded consequences. The requirement is that uncertainty enter the decision explicitly. The institution should know which possible prior outcomes make the next act permissible and which make it dangerous.

Permission to retry is conditional on the uncertainty that actually remains. A provider’s deduplication guarantee may depend on retaining the same identifier and material request, using the covered endpoint, and remaining within a stated time window. Those conditions must be established for the service concerned; the word “retry” establishes none of them. Where such guarantees are unavailable, another attempt needs a justification that addresses the possibility of an additional effect. Its authority may come from an applicable standing recovery mandate or from a fresh decision. Authorizing that next step does not resolve the earlier outcome. The original result must remain recorded as unknown until evidence resolves it.

This requires a distinction between the logical action and its technical attempts. Several transmissions can be attempts to perform one authorized operation; alternatively, what appears to be a retry can create a second commitment. Authority should attach to the appropriate unit. A system that grants fresh permission to every attempt may accidentally authorize accumulation beyond the original purpose. A system that prohibits every repeated transmission may prevent completion of an otherwise legitimate obligation. Recovery policy must address both errors.

There is also a limit to what compensation can repair. Withdrawing a notice may correct the formal record while leaving distress, disclosure, or a missed opportunity unresolved. Restoring a balance may not restore access to a service that was unavailable when needed. A recovery account should therefore include correction of state, communication with affected people, and any further remedy the situation requires. Treating recovery as a purely technical rollback would omit much of what makes the original action consequential.

Evidence itself can impose costs. Detailed records may expose sensitive information or create a new opportunity for misuse. Maintaining accountability does not require copying all private content into every audit system. An institution should retain the evidence needed for defined purposes, control access to it, and decide how long it remains justified to keep it. A reference to an authoritative record, a protected evidential store, or a limited commitment to material terms may be preferable to unrestricted duplication. The right arrangement depends on which claims must later be established.

The practical aim is an honest account that can support action and challenge. An interface should make uncertainty understandable without burying users in implementation details. An operator should be able to identify what remains unresolved and which actions are permitted while it is unresolved. An affected person should have a route to contest the institution's account. These are different audiences, and a useful evidence system must serve more than the team that built it.

8 The ethical basis of governed execution

The relationships examined in sections 5 and 7 matter because institutions act on people who have claims of their own. A claimant usually cannot inspect an office's internal operations before relying on its notice or suffering a deprivation. The office cannot discharge a duty to issue an authorized determination by showing an approval for one notice and a receipt for another. Nor can it discharge a duty to correct an error by repeating an operation without considering whether repetition creates a further harm. Where these duties are owed, they constrain how the institution delegates the work.

The argument has three premises. First, an institution can owe duties concerning the exercise of its powers independently of how it performs the work. Second, changing the means of performance does not by itself release the institution from those duties. Third, where the chosen means would predictably frustrate their discharge, the institution must adopt adequate safeguards, choose different means, or justify a change in the duties themselves. It follows that delegation requires proportionate assurance of the relationships those duties make necessary. Continuity of authority names that requirement.

The second premise needs defense. If an institution could extinguish an obligation merely by assigning its performance to another agent, the party bearing the obligation could unilaterally determine whether the other party retained its protection. The claimant's standing would depend on the office's internal division of labor. Where the duty is owed to the claimant, that is not a sufficient reason for release. A lawful and morally defensible transfer of responsibilities, a valid change in an agreement, or a justified revision of an institution can alter duties. Those changes require reasons beyond the fact that a different system now performs the task. Delegation of work and release from an obligation are separate matters.

An outcome-based objection challenges the first premise. If a system reliably produces better decisions, perhaps procedures have value only as instruments for achieving those results. On that view, requiring additional controls after their benefits are exhausted would be wasteful. I agree about redundant safeguards. I reject the assumption that a prediction score captures every relevant result. A procedure can protect a person's ability to introduce information, dispute the institution's jurisdiction, or challenge a rule that is being applied accurately. Those are interests in the terms of the relationship. A claimant can have a justified objection to an accurately applied rule, and an institution must be able to hear that objection rather than treating accuracy as its answer.

The obligation determines the assurance required. If evidence establishes that a standing procedure adequately protects the relevant interests, another approval layer needs its own justification. Aggregate reliability, however, cannot answer a dispute about whether this particular notice was authorized. The institution needs evidence suited to that claim. Delay, wrongful refusal, and the exposure created by record-keeping belong in the assessment alongside unauthorized action.

This argument places legitimacy before optimization in a specific sense. Before asking how efficiently an institution can pursue an objective, it must ask whether it may pursue that objective through the proposed means. Efficiency still matters. Delay, administrative expense, and unreliable service can themselves injure people. The claim is that these considerations belong within an account of justified action; they cannot supply all of that account. A more efficient wrongful deprivation remains wrongful.

Pettit's conception of freedom as non-domination gives a reason to reject benevolent performance as the whole standard (Pettit, 1997, Chapters 2 and 6). A power can remain arbitrary even when it is exercised favorably. The relevant constraint is that those subject to it need not depend on the holder's goodwill. Applied to AI-mediated institutions, this supports procedures through which affected interests can constrain decisions and challenges can make a practical difference. It does not entail that every person must consent to every act.

Records can support those constraints, but they can also preserve a perfectly documented injustice. A discriminatory mandate can be faithfully executed and impeccably recorded. In that case there may be no authorization failure in the institution's own terms; the defect is in its claim to legitimate power. Governed execution needs an outward-facing component if it is to contribute to justified governance: notice where appropriate, access to relevant reasons and evidence, and review capable of changing the result. Calling procedural conformity legitimate would erase the distinction on which this argument depends.

Dignity supplies a related argument. People affected by a decision have knowledge, claims, and objections of their own. A claimant may know that a record is wrong or that an apparently routine case includes an exceptional circumstance. A process that excludes all possibility of being heard can disregard that standing even if its average predictions are accurate. An adequate institutional procedure must give appropriate weight to the person's position as a participant in the relationship.

The legal requirements of due process vary with jurisdiction, institution, and type of decision. The moral argument for a fair opportunity to contest a consequential act is broader than any single legal formula. The two should be kept distinct. Articles 12 and 14 of the EU AI Act address record-keeping and human oversight for high-risk systems within the regulation's scope (European Union, 2024). Those provisions are relevant examples, not a general legal endorsement of the lifecycle proposed in this paper or a guarantee that using it satisfies the Act.

Professional and fiduciary relationships create further reasons for care. A person who entrusts an institution with sensitive interests may depend on duties that cannot be reduced to immediate instructions or measurable satisfaction. The precise legal duties differ among relationships. The general governance issue is whether delegation preserves the responsibilities that make the relationship acceptable. A system that follows a customer's expressed preference while violating a professional obligation has not solved the problem by being responsive.

Proportionality is necessary if these demands are to remain defensible. Requiring extensive review of every low-impact operation can waste attention and delay benefits. A more demanding procedure is justified when the stakes, uncertainty, exposure of others, or difficulty of remedy warrant it. Proportionality does not mean choosing a single risk score and allowing everything below a threshold. Some constraints are categorical; others depend on context; still others concern cumulative effects. The institution should be able to explain which kind of constraint it is applying.

Restraint also has moral costs. An assistant that declines to send a necessary warning, complete an authorized payment, or escalate an urgent problem may expose someone to avoidable harm. A governance system should account for duties to act as well as duties to refrain. Emergency arrangements should specify the conditions for exception, the discretion permitted, and the subsequent review. This is an institutional design requirement, not a claim that obedience is always morally obligatory. An act outside an existing mandate could be morally justified in an exceptional case, just as an authorized act could be wrong. The distinction requires examining the reasons for the exception; a system's assertion that its preferred action is helpful does not settle them.

Finally, authority over the controls requires scrutiny. Someone chooses the policy, decides who may change it, and determines which failures become visible. These choices distribute power. They should be open to appropriate review rather than concealed behind a claim that the software is merely enforcing rules. Governed execution is ethically defensible only if its controls remain connected to a defensible institution. Otherwise, it may make arbitrary power more orderly without making it less arbitrary.

9 Principles for institutional design

The argument yields six design principles. They make the institutional act the unit of assessment and connect safeguards to the obligations they serve.

First, define the act at the level at which the obligations arise. Several database updates may implement one decision; a sequence of modest disclosures may expose a sensitive record. Their combined effect must remain within the mandate.

Second, identify who can grant the authority and what discretion the grant permits. Completion of a task does not authorize a system to expand its own role. Additional authority must come from someone entitled to confer it.

Third, enforce the material terms across the relevant paths to effect. Changes after review, downstream tools, and concurrent operations can defeat a locally correct check. Test the actual paths, including administrative and emergency routes, and state where enforcement ends.

Fourth, fit human control to the judgment required. A standing mandate, an individual review, and an appeal serve different purposes. Each needs the authority, information, time, and practical options appropriate to its purpose. Evaluate those conditions under ordinary workloads.

Fifth, make reported status answerable to evidence. An approval, an attempt, and a confirmed effect establish different things. Connect the records needed to support the institutional claim, limit unnecessary collection of private information, and keep unresolved outcomes visible.

Finally, design recovery and contestability with the original action. Determine how the institution will investigate an uncertain outcome, justify another attempt, correct an error, and respond to a challenge. A plan for routine success leaves these responsibilities unanswered.

For the benefits office, these principles permit substantial autonomy. The assistant can draft and dispatch notices covered by a standing mandate, while referring material departures for review. The office must still establish which notice was issued, what the delivery evidence supports, and what further action is justified if the result remains uncertain. The claimant's route to challenge the decision must survive the automation of its delivery.

10 Governed execution as a field of inquiry

I propose governed execution as a field of inquiry because the exercise of a mandate presents connected problems that cross existing specialties. The material identity of an act matters to both its authorization and its evidence. The behavior of a remote service matters to whether another attempt is permissible. The institution's duties determine which controls and remedies are adequate. Studying these relationships together gives the field a defined subject: how institutional authority is preserved, evidenced, and challenged through human and machine operations.

That work requires collaboration among security engineers, domain professionals, auditors, and researchers in law, ethics, and human factors. A specialist must understand where those forms of expertise constrain one another. A technically feasible recovery procedure can still create an unauthorized commitment. A formally available appeal can still be useless if the evidence required to pursue it was never retained. These are reasons to organize inquiry around the whole act.

The research agenda follows from those problems. Which features of an act must remain fixed for an authorization to continue to apply? What evidence is sufficient to connect approval with performance? Under what conditions does an uncertain result warrant another attempt, a protective intervention, or a pause? How can people challenge both the institution's account of what happened and the legitimacy of the mandate itself? Answering these questions requires conceptual argument, system evaluation, and evidence about institutional practice.

The proposed five dimensions provide a starting point for empirical work. Researchers could ask independent reviewers to analyze the same set of incidents using this framework and established alternatives. The comparison should hold the available evidence constant and assess whether the framework improves the detection of failures, agreement among reviewers, or selection of remedies. Cases should include successful delegation, missing documentation, unauthorized but accurately recorded acts, and procedurally compliant injustice. A taxonomy that labels every problem after the fact but changes no diagnosis has limited value.

System studies could compare implementations under controlled changes to recipients, material content, mandate expiry, and shared limits. Recovery studies should include lost acknowledgments and providers with different deduplication guarantees. Human-factors research could test whether review interfaces help people detect material deviations under realistic workloads. Outcomes should include false refusals, delay, unnecessary disclosure, and unresolved cases, as well as unauthorized effects. A control can improve one measure while making the service worse for the people it is meant to protect.

These studies would need strong baselines. Comparing a carefully designed governed-execution system with an agent given unrestricted access would demonstrate little about the framework's distinctive contribution. More informative comparisons would use mature access control, transactional safeguards, and established accountability procedures. Recent work on authenticated delegation, enterprise control, and reviewable decisions supplies relevant alternatives. The issue is whether the proposed integration adds explanatory or practical value once those alternatives are taken seriously.

A separate field must earn its place through explanatory and practical gains. Comparative work should test whether studying these relationships together reveals a material mismatch, resolves a disagreement about a remedy, or improves the design of institutional safeguards. If existing fields accomplish the same work without a separate specialization, the case for disciplinary independence weakens. The obligations remain. I defend the field proposal because these connections deserve sustained inquiry, and its value should be judged by how much that inquiry improves institutions' ability to answer for their actions.

11 Limitations and objections

This paper presents a conceptual argument. It reports no experiment, deployment evaluation, or systematic review of the whole literature. The five dimensions and lifecycle are analytical proposals. Their usefulness, completeness, and costs remain to be assessed. The examples illustrate distinctions; they do not establish how often particular failures occur. Nor does the paper establish that authority failures are more prevalent or more important than other problems in AI deployment.

One objection is that continuity of authority restates ordinary accountable delegation. The premise that duties survive delegation is familiar. The contribution here lies in specifying what follows for the relationship between a mandate and its exercise. Section 5 separates valid component behavior from an authorized institutional act; section 7 shows why an unchanged initial mandate can require different recovery decisions under different possible effects. Section 8 grounds the significance of those distinctions in obligations to affected people. The account uses established technical ideas to specify a philosophical problem and its practical consequences. A system that already preserves the required relationships satisfies the account; it need not be rebuilt or renamed.

A second objection concerns regress. If an authorizer needs authorization and a control needs another control, the proposal might demand an endless hierarchy. It does not. Technical verification ends at institutional sources of standing, such as the rules establishing an office or an agreement constituting a role. Those sources remain open to legal and moral criticism; another credential cannot resolve that criticism. Continuity requires an inspectable allocation of responsibility under stated assumptions, not a final machine-issued certificate of the institution's legitimacy.

A fully reliable system with an appropriately justified mandate is a limiting case, not a counterexample. If it supplies all the assurance the relevant obligations require, the argument warrants no additional external control merely because the system is artificial. In practice, asserting that reliability does not establish it. How much independent evidence is needed depends on the consequences of relying on a mistaken claim. The paper offers no general rule that a model must always be checked by a separate deterministic component.

Another objection concerns the instability of institutional meaning. Rules can conflict, discretion may be irreducible, and people may disagree reasonably about an action's status. Governed execution cannot eliminate these features by converting them into machine-readable fields. It should make unresolved questions visible, preserve the grounds for judgments that are made, and support referral to an appropriate decision-maker. A representation that conceals disagreement behind a precise schema would defeat that purpose.

The argument also leaves open difficult questions about future machine standing. It does not establish that only biological persons could ever hold authority. Its immediate concern is the present practice of deploying systems through human institutions and claims of delegated power. If a society recognized an AI system as an office-holder or legal person, the source, limits, and justification of that status would still require examination. Greater capability would remain insufficient, by itself, to answer those questions.

There are limits to action-centered analysis. Harms can arise from patterns of individually authorized acts, from a policy's distributional effects, or from an institution's failure to act at all. The framework therefore needs to be connected to scrutiny of policies, aggregate outcomes, and duties of service. It cannot replace them. A well-governed dispatch process will not make an unjust eligibility rule defensible.

Finally, legal requirements and professional duties vary. A general philosophical account cannot determine the legality of a particular deployment or define universal classes of nondelegable acts. The same is true of evidential sufficiency: what justifies one low-impact operation may be inadequate for a deprivation of rights. The proposed practice requires domain knowledge and judgment at precisely these points. The framework organizes that work; it cannot replace the judgment the work requires.

12 Conclusion

An institution cannot establish that an action was authorized merely by showing that each component worked as designed. It must connect the mandate to the materially relevant act and support its account of what followed. The approved notice and the dispatched notice must correspond in the respects that matter to the authorization. When the outcome remains unknown, the institution must justify its next step against the possibility that the first already took effect.

This relational account gives the authority gap a precise object of inquiry. It separates a missing mandate from an exceeded mandate, an unauthorized act from an evidential failure, and a further attempt at one act from an additional institutional effect. Those distinctions change what needs to be repaired and what can be done while uncertainty remains. Continuity of authority explains why institutions owe attention to those relationships: changing how they act does not, by itself, release them from obligations to the people affected.

Governed execution brings that work into a common field of inquiry. It supports standing mandates and substantial autonomy wherever the required relationships hold. It also requires institutions to preserve meaningful challenge and remedy, including challenges to rules that their systems execute correctly. The institution must answer for the act it actually performs. The claimant receives a notice, and it is for that notice that the office must answer.

Declarations

Competing interest. I am the founder of Aetheria Intelligence, which advertises AI governance advisory services and is developing technology for delegated authority and execution governance. The company has a commercial interest in the subject of this article. This article develops a philosophical argument and does not evaluate or validate the company's products.

AI assistance. ChatGPT (OpenAI) was used in editorial revision, development of objections and responses, and assistance with literature searches and citation checking. Claude (Anthropic) provided an additional review of the argument, contribution, structure, references, and formatting. I take full responsibility for the positions advanced and for the accuracy and integrity of the final manuscript.

Data availability. This conceptual article reports no original empirical data. Its constructed examples illustrate the argument and are not presented as observed cases.

References

American Law Institute. (2006). Restatement of the law third, agency. American Law Institute. https://www.ali.org/publications/restatement-law-third/agency

Austin, J. L. (1962). How to do things with words. Clarendon Press.

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Clark, D. D., & Wilson, D. R. (1987). A comparison of commercial and military computer security policies. 1987 IEEE Symposium on Security and Privacy. https://doi.org/10.1109/SP.1987.10001

Cobbe, J., & Singh, J. (2020). Reviewable automated decision-making. Computer Law & Security Review, 39, 105475. https://doi.org/10.1016/j.clsr.2020.105475

Davidovic, J. (2023). On the purpose of meaningful human control of AI. Frontiers in Big Data, 5, 1017677. https://doi.org/10.3389/fdata.2022.1017677

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems, 37, 82895–82920. https://doi.org/10.52202/079017-2636

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

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

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

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. Advances in Neural Information Processing Systems, 29. https://proceedings.neurips.cc/paper_files/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html

Hu, V. C., Ferraiolo, D., Kuhn, R., Schnitzer, A., Sandlin, K., Miller, R., & Scarfone, K. (2014). Guide to attribute based access control (ABAC) definition and considerations (NIST SP 800-162; updated August 2019). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-162

ISO/IEC. (2023). Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023). International Organization for Standardization. https://www.iso.org/standard/42001

Jørgensen, B. N., & Ma, Z. G. (2026). Governing agentic AI in enterprise workflows: A bounded-autonomy framework for delegated authority and controlled execution. Information, 17(9), 923. https://doi.org/10.3390/info17090923

Kolt, N. (2026). Governing AI agents. Notre Dame Law Review, 101, 335. https://ndlawreview.org/governing-ai-agents/

Laux, J., & Ruschemeier, H. (2025). Automation bias in the AI Act: On the legal implications of attempting to de-bias human oversight of AI. European Journal of Risk Regulation, 16(4), 1519–1534. https://doi.org/10.1017/err.2025.10033

Lifshitz, L. R., & Hung, R. (2024). BC Tribunal confirms companies remain liable for information provided by AI chatbot. Business Law Today. American Bar Association. https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/

Lodderstedt, T., Richer, J., & Campbell, B. (2023). OAuth 2.0 rich authorization requests (RFC 9396). Internet Engineering Task Force. https://doi.org/10.17487/RFC9396

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

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

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

OWASP Gen AI Security Project. (2026, September 1). Agent Control Standard (ACS). https://genai.owasp.org/resource/agent-control-standard-acs/

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

Pettit, P. (1997). Republicanism: A theory of freedom and government. Oxford University Press.

Raz, J. (1986). The morality of freedom. Clarendon Press.

Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. AI & Society, 41, 259–278. https://doi.org/10.1007/s00146-025-02422-7

Royal Commission into the Robodebt Scheme. (2023). Report of the Royal Commission into the Robodebt Scheme. Australian Government. https://robodebt.royalcommission.gov.au/publications/report

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, 15. https://doi.org/10.3389/frobt.2018.00015

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36, 68539–68551. https://doi.org/10.52202/075280-2997

Searle, J. R. (1995). The construction of social reality. Free Press.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. AAAI Workshop on AI and Ethics. https://intelligence.org/files/Corrigibility.pdf

South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Chan, A., & Pentland, A. (2025). Position: AI agents need authenticated delegation. Proceedings of Machine Learning Research, 267, 82211–82231. https://proceedings.mlr.press/v267/south25a.html

Thurlow, R. (2009). RPC: Remote procedure call protocol specification version 2 (RFC 5531). Internet Engineering Task Force. https://doi.org/10.17487/RFC5531

United States. (2000). Electronic Signatures in Global and National Commerce Act, 15 U.S.C. § 7001(h). https://www.law.cornell.edu/uscode/text/15/7001

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8, 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Weber, M. (1978). Economy and society: An outline of interpretive sociology (G. Roth & C. Wittich, Eds.). University of California Press. (Original work published 1922)

Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., & Press, O. (2024). SWE-agent: Agent-computer interfaces enable automated software engineering. Advances in Neural Information Processing Systems, 37, 50528–50652. https://doi.org/10.52202/079017-1601

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations. https://arxiv.org/abs/2210.03629