

Does Human Oversight of AI-Assisted Decisions Work? A Systematic Review of the Evidence and a Framework for Meaningful Human Control

Abid Iqbal

abidiqbal.apex@gmail.com

Iqra University <https://orcid.org/0000-0002-1897-0393>

Saiqa Ameen

Sindh Agriculture University

Jamil Channa

Shah Abdul Latif University

Imran Maqbool Thebo

University of Electronic Science and Technology of China

Systematic Review

Keywords: meaningful human control, human oversight, AI-assisted decision-making, AI governance, algorithmic accountability, systematic review

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11445281/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Does Human Oversight of AI-Assisted Decisions Work? A Systematic Review of the Evidence and a Framework for Meaningful Human Control

Abstract

Human oversight has become the default safeguard for AI-assisted decisions: the EU Artificial Intelligence Act, data protection law and many organizational policies require a human to review or override algorithmic outputs. Whether such oversight gives decision makers meaningful control, and whether it improves decisions, is far less clear. We conducted a PRISMA 2020 systematic review of 186 studies from Scopus and Web of Science on the governance of AI-assisted decision-making and graded each study by the strength of evidence its design can provide. Governance was conceptualized mainly through accountability (96% of studies) and transparency (94%), and human-centered control appeared in 70% of studies, yet its measurement was fragmented and human control was rarely measured as a single construct. Although 50 studies reported positive relationships between governance and outcomes, only three (1.6%) used designs able to support causal claims; both causal studies measured perceived legitimacy rather than decision quality, and neither isolated override, calibrated reliance or contestability. The evidence for human oversight is therefore largely normative and associational. We decompose meaningful human control into five measurable components (information and understanding, decision authority, capacity to override, competence and timing, and answerability), propose an eight-component framework with seven testable propositions, and set out a research agenda for testing whether oversight works in practice.

Keywords: meaningful human control; human oversight; AI-assisted decision-making; AI governance; algorithmic accountability; systematic review

1. Introduction

AI now operates inside decisions that directly affect individuals, organizations and institutions, from welfare fraud detection, child protection and policing to credit approval and clinical care (Veale et al., 2018; Wilson & van der Velden, 2022; Waldman & Martin, 2022; Wieringa, 2023; Hillo et al., 2025; Lin et al., 2025; Liu, 2025). Because such systems classify, predict, recommend and prioritize (Hauer et al., 2023; Papagiannidis et al., 2025), governing them cannot be reduced to securing model accuracy. The decisive questions are institutional: who is responsible for a decision, what information reaches the decision maker, what authority a human reviewer actually holds, and how an affected person can seek review and redress (Hunt & McKelvey, 2019; Koniakou, 2021; Hickok, 2022; Lazcoz & De Hert, 2023). The prevailing regulatory answer is human oversight. Article 14 of the EU Artificial Intelligence Act requires effective human oversight of high-risk systems, and data protection law gives individuals rights in relation to decisions based solely on automated processing (Lazcoz & De Hert, 2023; European Parliament & Council of the European Union, 2024). Critics warn, however, that oversight requirements can legitimize flawed systems without improving the decisions they produce (Green, 2022). Whether human oversight works, and what would have to be true for it to give people meaningful control, is the question this review addresses.

1.1 The problem: governance concepts without governance constructs

AI governance refers to the structures, mechanisms, practices and institutional arrangements that guide, constrain, review and monitor AI decision-making and render it accountable (Veale et al., 2018; Hunt & McKelvey, 2019). It may encompass accountability, transparency, human oversight, fairness, risk management, regulation, privacy, participation and contestability (Koniakou, 2021; Hickok, 2022; Wilson & van der Velden, 2022; Hauer et al., 2023). Yet naming a concept does not make it a governance construct. We therefore retained only those concepts that discipline the decision system and shape its institutional use and allocation of responsibility (Lazcoz & De Hert, 2023; Papagiannidis et al., 2025).

1.2 The human gap

Meaningful human control is not secured merely by appointing a reviewer (Koulu, 2020; Green, 2022). The reviewer must grasp the relevant information and the limits of the AI (Cornelissen et al., 2022; Lazcoz & De Hert, 2023), and that person, or a human team, must hold genuine authority to suspend, modify or overturn a decision (Koulu, 2020; Green, 2022; Torrecilla-Pinero, 2026). Several studies nonetheless document oversight that is merely nominal, in which the human role is reduced to endorsing an automated recommendation (Lazcoz & De Hert, 2023; Torrecilla-Pinero, 2026). Policy thus risks mandating the form of oversight while leaving its function unmeasured.

1.3 Research questions and contributions

- **RQ1.** How has AI governance been conceptualized in AI decision-making?
- **RQ2.** How has AI governance been operationalized and measured?
- **RQ3.** What evidence exists on AI governance and AI decision-making outcomes, and how strong is it?
- **RQ4.** What theoretical, methodological and practical gaps remain?

This review makes four contributions. **First**, it examines the evidential basis of human oversight, grading 186 studies by the strength of their designs and separating the direction of a reported relationship from the identification strength of the design: only 1.6% of the corpus can support causal inference, and none of the causal evidence concerns decision quality. **Second**, it reframes governance from the presence of mechanisms to the exercise of control, decomposing meaningful human control into components that can be observed and measured (Koulu, 2020; Cornelissen et al., 2022; Green, 2022; Lazcoz & De Hert, 2023). **Third**, it proposes an eight-component, human-centered framework with seven testable propositions that links governance architecture to decision outcomes through meaningful human control (Gordon et al., 2022; Hickok, 2022; Lahusen et al., 2024; Papagiannidis et al., 2025). **Fourth**, it converts the identified gaps into a prioritized, design-specific research agenda. To our knowledge, no earlier review combines these four elements for AI-assisted decision-making specifically; Section 5.4 situates the work against the closest frameworks.

1.4 Contribution relative to prior reviews

Two earlier reviews are closest to ours. Zuiderwijk et al. (2021) reviewed 26 studies on the implications of AI use in public governance and found the literature to be mostly exploratory, conceptual and qualitative, calling for more empirical and explanatory research. Papagiannidis et al. (2025) reviewed 48 articles on responsible AI governance in organizations and proposed structural, procedural and relational governance practices. In parallel, the philosophical literature has developed meaningful human control as a normative standard, from the tracking and tracing conditions of Santoni de Sio and van den Hoven (2018) to design

proposals that support reflective human judgment in decision support (Cornelissen et al., 2022). The present review connects these strands in three ways. First, it examines the governance of AI-assisted decisions rather than AI governance or AI use in general. Second, it grades all 186 included studies by the strength of evidence their designs can provide: only three studies can support causal claims, and both causal studies measure perceived legitimacy rather than decision quality. Third, it translates meaningful human control from a normative standard into measurable components and testable propositions, so that whether oversight works can be studied empirically.

2. Conceptual background

Governance of AI is not a single value but a multidimensional, multilevel process. Ethics, trust, explainability and adoption qualify as governance features only when they are tied to institutional responsibility, control, monitoring or rules within the decision process (Hunt & McKelvey, 2019; Koniakou, 2021; Gordon et al., 2022; Hickok, 2022; Lazcoz & De Hert, 2023; Lahusen et al., 2024; Papagiannidis et al., 2025).

2.1 Governance dimensions

Accountability and auditability emphasize traceability and obligation (Hunt & McKelvey, 2019; Koniakou, 2021), whereas transparency and explainability help people understand outputs and their limits (Gordon et al., 2022; Hickok, 2022). Human involvement concerns supervision and intervention; fairness and bias mitigation address inequitable outcomes (Lazcoz & De Hert, 2023; Lahusen et al., 2024). Risk management anticipates harm and uncertainty, regulation imports legal and policy settings, and privacy and data governance safeguard data (Koniakou, 2021; Papagiannidis et al., 2025). Further dimensions include participation, contestability, independence, documentation and administrative arrangements (Hickok, 2022; Lazcoz & De Hert, 2023; Lahusen et al., 2024).

2.2 Meaningful human control and calibrated reliance

Meaningful human control rests on five interlinked elements: information and understanding, decision authority, the practical capacity to intervene or override, competence and timing, and answerability (Koulu, 2020; Cornelissen et al., 2022; Green, 2022; Lazcoz & De Hert, 2023; Torrecilla-Pinero, 2026). A human in the loop does not by itself establish control; the operative question is how much authority that person can actually exercise. Control also demands calibrated reliance. Automation bias can arise, but rejecting most AI recommendations discards useful information (Koulu, 2020; Cornelissen et al., 2022). Decision makers therefore need both the capability and the authority to accept or reject a recommendation after weighing the quality of the information and the degree of uncertainty (Green, 2022; Lazcoz & De Hert, 2023; Torrecilla-Pinero, 2026).

2.3 Contestability, organizational enablement and context

A decision is contestable when it can be reviewed, challenged or appealed and a remedy pursued (Lyons et al., 2021; Costanza-Chock et al., 2022; Waldman & Martin, 2022; Wieringa, 2023). Organizational enablement (rules, procedures, training, resources, role clarity, technical support and escalation channels) translates policy into practice (Pi, 2024). Governance requirements also scale with risk, uncertainty, task, technical skills and organization type (Watkins et al., 2021; Green, 2022; Alfrink et al., 2023); decisions that are high-risk and hard to reverse call for stronger controls (Fan, 2025; Lin et al., 2025; Nouis et al., 2025; Workum et al., 2025; Erdemir & Sumbas, 2026; Kim & Shin, 2026; Tekin & Stoycheva, 2026; Widlak, 2026).

3. Methods

3.1 Design and reporting

The review proceeded in two phases: an exploratory and analytical phase, followed by a formal rebuild on 29 August 2026 in which search, eligibility, screening, extraction, appraisal and synthesis were organized in a design-sensitive sequence. Reporting follows PRISMA 2020 (Page et al., 2021a, 2021b). Appraisal was matched to study design using CASP, JBI, RoB 2, ROBINS-I, MMAT and a structured appraisal. Quality and eligibility were treated separately so that weaker evidence could be reported transparently rather than silently excluded. The review was not prospectively registered in PROSPERO or OSF.

3.2 Search

The formal search identified 848 records (722 from Scopus; 126 from Web of Science); removing 108 duplicate clusters left 740 unique records. The search was restricted to English-language Articles, Reviews and Conference or Proceedings Papers published between 1 January 2000 and 29 August 2026, with no subject-area restriction. Three concept blocks (AI, AI governance, and AI-assisted or algorithmic decision-making) were adapted to the syntax of each database; the full Boolean strings appear in Table S1.

Disclosure of the exploratory search. An earlier exploratory export (around 11 August 2026) yielded 225 records (152 from Scopus; 73 from Web of Science), reduced to 212 after de-duplication. Eighteen were excluded at title and abstract, 194 proceeded to full-text retrieval, and 104 were retained. Because this retrieval trail could not be fully reconstructed, the review was rebuilt around a new search, and the 104-study historical corpus was not carried into the PRISMA flow or the denominator of 186.

3.3 Eligibility and screening

Studies were eligible if they treated AI governance as a practical construct, mechanism, framework, process or institutional arrangement and meaningfully linked it to AI decision-making. A passing mention of ethics, trust, explainability, adoption, fairness, privacy or technical performance, without a practical connection to governance, was insufficient. Theoretical, review and empirical studies were classified by evidential role.

Titles and abstracts were first triaged with ChatGPT (OpenAI), which assigned the 740 records to Include ($n = 468$), Unclear ($n = 128$) or Exclude ($n = 144$). These labels were advisory only: the human team reviewed every record, and human judgment overrode the tool whenever they disagreed. In a post hoc consistency check, with Unclear and Exclude collapsed into Not Include, agreement with the final human decisions was 88.4% (Cohen's $\kappa = 0.761$). This statistic captures consistency between the AI triage and the human decisions; it is not inter-rater reliability between independent human reviewers.

3.4 Final corpus

Of the 740 records, 406 advanced to full-text retrieval and 334 were excluded. Of these 406, 153 could not be retrieved, leaving 253 for assessment; 58 were then excluded, leaving 195. Study-level reconciliation removed six residual duplicates and three out-of-scope studies (Table S9), yielding a final corpus of 186 unique studies (Figure 1; the full list is in Table S3, and the screening and eligibility logs are in Tables S5 and S6).

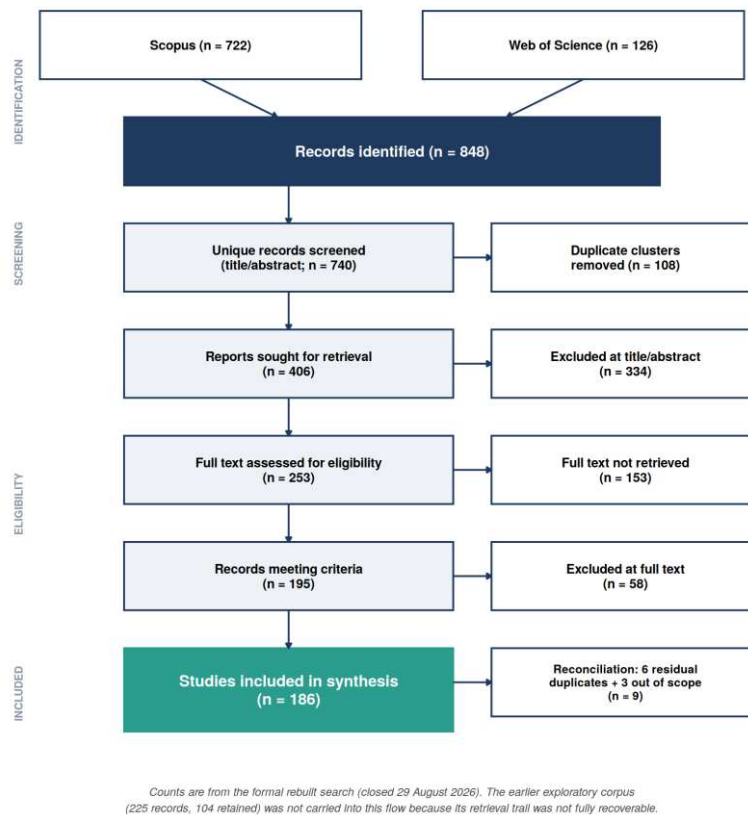


Figure 1. PRISMA 2020 flow diagram of the formal rebuilt search and study selection. Created with the assistance of ChatGPT (OpenAI) and Claude (Anthropic) and reviewed by the authors.

3.5 Extraction and appraisal

We extracted bibliographic data, study design, theory, method, sample and findings, together with definitions of AI governance and its dimensions, operationalization and measurement details, RQ3 results, statistical relationships and strength of evidence. Causal evidence was coded separately from associational and descriptive evidence. Where measurement items, reliability and validity data or page locators were absent, we made no inferences. Appraisal tools were assigned by design, whereas evidence classification asked a different question: did the study identify or estimate the effect of a governance mechanism on a decision outcome? Consequently, some experimental or intervention studies received RoB 2 or ROBINS-I appraisal without being classified as causal evidence for the governance–outcome relationship.

RoB 2 (Sterne et al., 2019) was applied to randomized trials, ROBINS-I (Sterne et al., 2016) to non-randomized intervention studies, MMAT (Hong et al., 2019) to mixed-methods studies, CASP to qualitative studies and JBI tools to reviews and observational or analytical studies. Computational, design-science and conceptual studies were assessed with a structured appraisal of methodological transparency, construct or design coherence, validation or evidence grounding, robustness, reproducibility, assumptions and stated limitations. Within the named-tool rows, N/A indicates that the information reported was insufficient for a defensible judgment; in the design-sensitive row, N/A indicates that conventional risk-of-bias criteria did not apply. Ratings were not used to exclude studies; study-level ratings and rationales are in Table S8.

3.6 Synthesis

We used a structured narrative synthesis. For RQ2, studies were compared across families of operationalization. For RQ3, positive, negative, mixed, descriptive or associational, and conceptual or proposed relationships were coded separately, and causal evidence was kept distinct from association. Heterogeneity in constructs, designs, outcomes and effect metrics precluded meta-analysis. The coding framework and extraction matrix are provided in Tables S4 and S7. Figures 2, 3 and 5 were produced in Python using code generated with the assistance of ChatGPT (OpenAI) and Claude (Anthropic). The code was run on the authors' coded extraction data (Table S7), and all plotted values were checked by the authors against Tables 1–5.

4. Results

4.1 RQ1: How has AI governance been conceptualized?

Across the 186 studies, AI governance emerged as a multidimensional, multilevel arrangement (Table 1; Figure 2a). Accountability (96%) and transparency or explainability (94%) were near-universal, followed by rights, ethics and values (80%), regulation (74%) and human-centered control (70%). By contrast, socio-technical or multilevel framings appeared in only nine studies (5%), despite the field's rhetoric of multilevel governance. Codes were applied only where a dimension received substantive treatment, and families overlap.

Table 1. Conceptual families of AI governance (% = n/186; overlapping codes).

#	Governance family	n	%
1	Accountability-centered	178	96%
2	Transparency / explainability	174	94%
3	Rights, ethics or values	148	80%
4	Regulatory / legal	137	74%
5	Human-centered or control	130	70%
6	Risk and safety	101	54%
7	Participatory / stakeholder	91	49%
8	Organizational / institutional	71	38%
9	Trust or legitimacy	61	33%
10	Contestability / procedural justice	61	33%
11	Life-cycle or adaptive	51	27%
12	Socio-technical / multilevel	9	5%

Consensus on responsibility and accountability coexists with divergent terminology and analysis (Koulu, 2020; Cornelissen et al., 2022; Green, 2022). The pivotal distinction lies between formal human management and real human control: a rule or policy is effective only when it is paired with adequate information, genuine authority and the power to intervene (Lazcoz & De Hert, 2023; Torrecilla-Pinero, 2026).

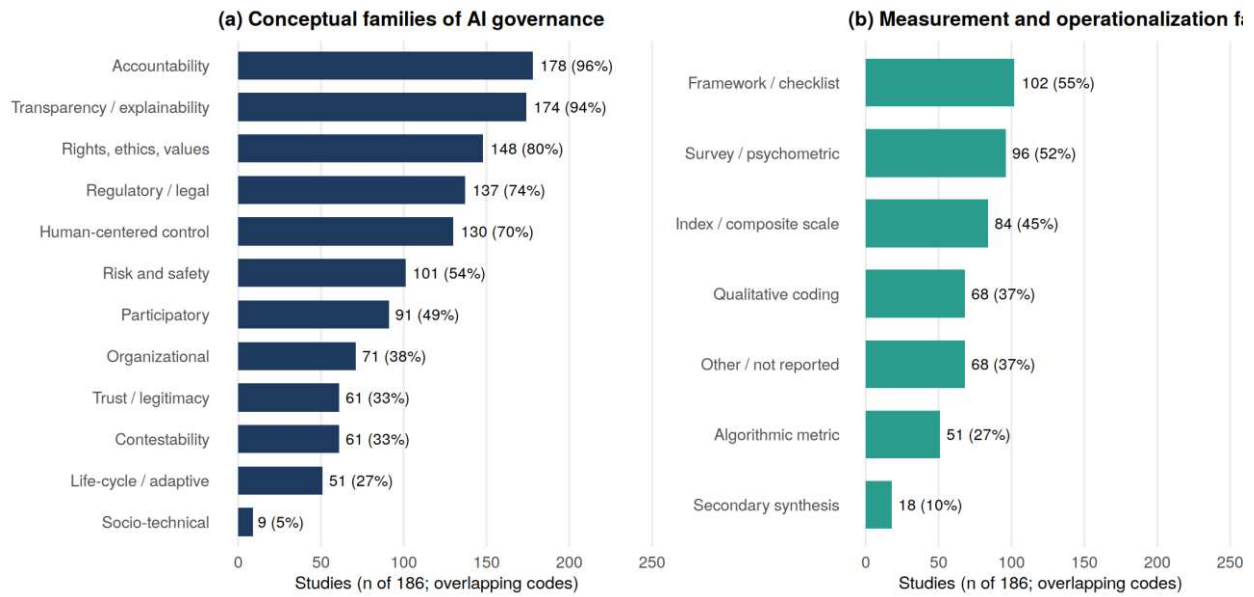


Figure 2. Prevalence of (a) conceptual families of AI governance and (b) measurement families across the 186 included studies. Codes overlap.

4.2 RQ2: How has AI governance been operationalized and measured?

No single approach to measurement has emerged (Table 2; Figure 2b). Frameworks, checklists and protocols (55%), surveys and psychometric instruments (52%) and indices or composite scales (45%) were most common, followed by qualitative coding (37%) and algorithmic or performance metrics (27%). Notably, 37% of studies fell into the “other or not reported” category, which signals under-specified operationalization. Categories are paper-level and overlapping; they therefore do not estimate the prevalence of any single instrument.

Table 2. Governance measurement and operationalization families (% = n/186; overlapping codes).

#	Measurement family	n	%
1	Framework, checklist or protocol	102	55%
2	Survey or psychometric	96	52%
3	Index or composite scale	84	45%
4	Qualitative or thematic coding	68	37%
5	Other or not reported	68	37%
6	Algorithmic or performance metric	51	27%
7	Secondary synthesis or review coding	18	10%

Measurement gap. Fragmentation impedes comparison because constructs, units of analysis, respondents and scoring shift from study to study (Veale et al., 2018; Hauer et al., 2023; Hulstijn, 2023; Lin et al., 2025; Natali & Cabitza, 2025). Studies also reported items, reliability, validity and source locators inconsistently, and the missing information could not be recovered. Future work should explicitly test construct definition, item provenance, dimensional structure, reliability, validity and measurement invariance (Papagiannidis et al., 2025; Ali et al., 2026; Kim & Shin, 2026; Zieliński et al., 2026).

Human-control measurement gap. Most of the literature treats oversight either as approval or as a human in the loop (Koulu, 2020; Cornelissen et al., 2022; Green, 2022). Information, authority, override, competence, timing and accountability are consequently not modeled as a single construct (Lazcoz & De

Hert, 2023; Torrecilla-Pinero, 2026). Developing a multidimensional measure of meaningful human control is a priority for follow-up research.

4.3 RQ3: What is the evidence on governance and decision outcomes?

The 186 studies comprised 50 positive, 18 negative, 30 mixed, 37 descriptive or associational and 51 conceptual or proposed relationships (Table 3; Figure 3b). Direction must not be mistaken for causal effect: although 27% of studies report a positive relationship, this describes an association, not what governance actually does.

Table 3. Relationship profile (% = n/186).

#	Relationship	n	%	Interpretation
1	Positive	50	27%	Positive association or effect; not necessarily causal.
2	Negative	18	10%	Adverse association or effect, or trade-off.
3	Mixed	30	16%	Direction varies by outcome or condition.
4	Descriptive / associational	37	20%	Observed association without causal identification.
5	Conceptual / proposed	51	27%	Mechanism proposed but untested.

4.3.1 The causal evidence deficit

By level of evidence, the corpus comprised 68 conceptual, 50 associational and 34 descriptive studies, 31 review syntheses, two causal studies and one causal-capable study (Table 4; Figure 3a). Only three of the 186 studies (1.6%) therefore support causal or causal-capable inference. Figure 4 makes the gap explicit: 178 studies address accountability and 130 address human control, yet just three can show whether such arrangements change decisions.

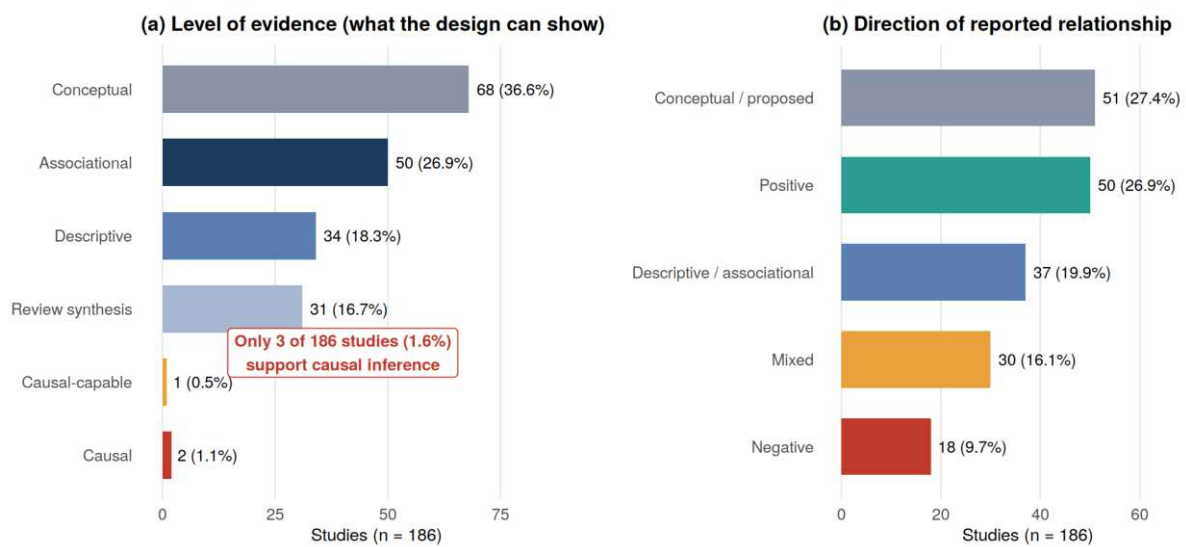


Figure 3. Evidence profile of the 186 studies: (a) level of evidence by design; (b) direction of the reported governance–outcome relationship. Direction does not imply causation.

Table 4. Causal evidence profile (% = n/186).

#	Evidence level	n	%	Definition
1	Conceptual	68	37%	Develops or explains concepts, frameworks or propositions without empirical tests.
2	Associational	50	27%	Examines associations between constructs; no causal claim.
3	Descriptive	34	18%	Describes practices, perceptions or conditions; no relationship test.

#	Evidence level	n	%	Definition
4	Review synthesis	31	17%	Synthesizes the literature; not primary empirical research.
5	Causal	2	1%	Design can identify or estimate a causal effect.
6	Causal-capable	1	0.5%	Design has causal potential, but the analysis was inadequate for full classification.

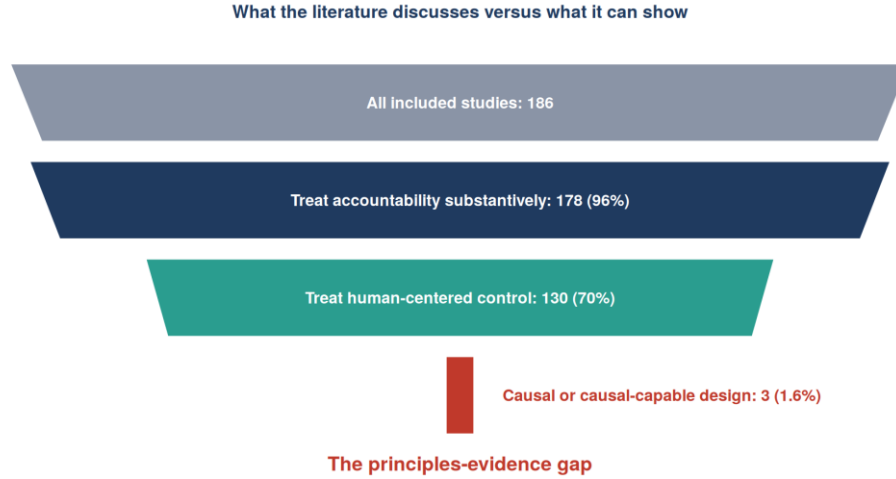


Figure 4. The principles–evidence gap: breadth of conceptual treatment versus the number of studies able to support causal inference. Created with the assistance of ChatGPT (OpenAI) and Claude (Anthropic) and reviewed by the authors.

Both causal studies were randomized experiments. Hillo et al. (2025) conducted a 2×2 factorial survey experiment with Finnish top-level public administrators and citizens in an automated child-protection decision context, manipulating algorithmic transparency or disclosure and human discretion; outcomes included perceived legitimacy, openness, responsibility and trustworthiness. Waldman and Martin (2022) used a randomized factorial vignette survey spanning algorithmic government decisions (education funding, policing, healthcare, parole and pothole repair), varying governance regimes, including human-in-the-loop oversight, with perceived legitimacy as the principal outcome. Liu (2025) compared governance-oriented transparency approaches in a double-blind design, including an AI-assisted credit-approval case and medical applications; outcomes included explanation consistency, user understanding, regulatory review time and throughput. We classified it as causal-capable rather than causal because it contrasts a LIME baseline with a proposed SHAP-GRU framework without random assignment to governance conditions or a design that isolates the effect of a governance mechanism at the decision level.

Two features of this causal base deserve emphasis. Both causal studies measure *perceived legitimacy*, a perception-based outcome, rather than decision accuracy, error correction or harm avoidance, and none isolates override, calibrated reliance or contestability as the manipulated mechanism. The strongest evidence is thus also the narrowest.

4.3.2 Outcomes, trade-offs and boundary conditions

Reported outcomes included trust and acceptance, accountability, transparency, fairness, safety, risk and efficiency, alongside human agency, privacy, compliance, participation, contestability and organizational outcomes (Veale et al., 2018; Green, 2022; Hesselmann, 2024; Lukkien et al., 2024; Yousefi, 2024; Devereaux, 2025; Clavijo et al., 2026; Johari, 2026; Pan et al., 2026; Vigeant et al., 2026). Studies also reported trade-offs, including added workload, time investment, friction and procedural burden, so the benefits of governance cannot be judged by any single outcome. Risk, uncertainty, task complexity, workload, expertise, organizational capability, regulatory context and human–AI configuration recurred as

boundary conditions (Watkins et al., 2021; Green, 2022; Alfrink et al., 2023; Fan, 2025; Lin et al., 2025; Nouis et al., 2025; Workum et al., 2025; Erdemir & Sumbas, 2026; Kim & Shin, 2026; Tekin & Stoycheva, 2026; Widłak, 2026). Governance outcomes thus depend on context as well as on mechanism, and no single arrangement suits every decision setting.

4.4 Quality appraisal

Of the 186 studies, 13 were rated high quality, 80 moderate and 93 N/A (not rateable or not applicable); none was rated low (Table 5; Figure 5). The large N/A share (50%) reflects the conceptual and design-science character of much of the corpus and, within named tools, insufficient reporting. The absence of any low rating should be read in this light: ratings discriminate among rateable studies, not across the whole corpus.

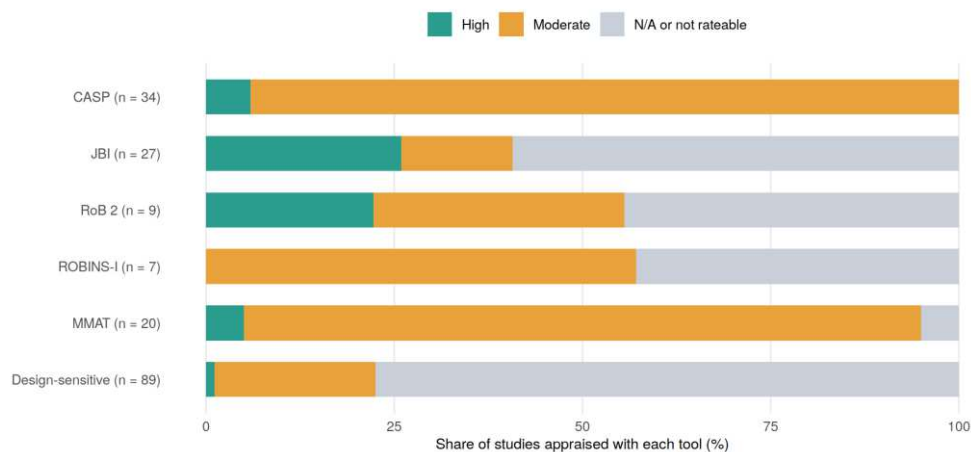


Figure 5. Design-sensitive quality appraisal by tool (percentage of studies appraised with each tool).

Table 5. Design-sensitive quality appraisal mapping (186 studies).

#	Appraisal mapping	Studies	High	Moderate	N/A
1	CASP	34	2	32	0
2	JBI	27	7	4	16
3	RoB 2	9	2	3	4
4	ROBINS-I	7	0	4	3
5	MMAT	20	1	18	1
6	Design-sensitive / N/A	89	1	19	69
	Total	186	13	80	93

4.5 RQ4: Theoretical, methodological and practical gaps

Theoretical gaps concern theory integration, meaningful human control and distributed accountability. *Methodological* gaps include the lack of measurement convergence and causal identification, and thin evidence on override effectiveness, automation bias and risk sensitivity. *Practical* gaps involve workload, contestability and implementation fidelity. In short, the literature is abundant in principles but sparse in evidence about mechanisms, their link to outcomes and the measurement of implementation.

5. A human-centered AI governance framework

5.1 Framework

Figure 6 integrates governance arrangements with the conditions under which human decision makers can exercise meaningful control over AI-assisted decisions; Table S2 traces each component to the studies that support it, and Table 6 defines the eight components. The pathway is moderated by risk, uncertainty, reversibility, task complexity, workload and institutional context. It is an evidence-informed integrative framework, not a validated causal model.

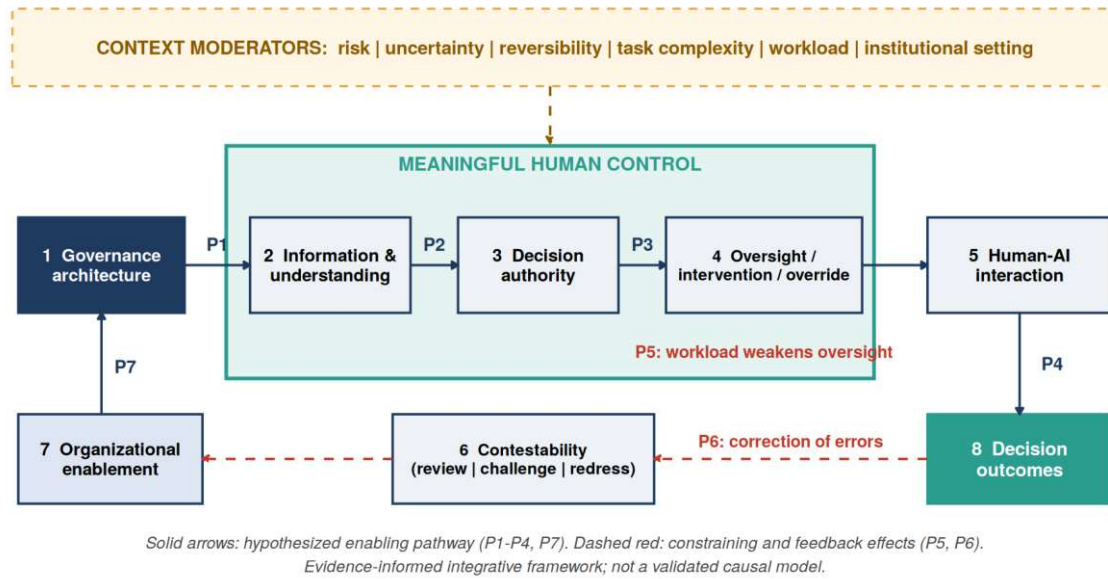


Figure 6. Human-centered AI governance framework: eight components, meaningful human control, context moderators and seven testable propositions (P1–P7). Created with the assistance of ChatGPT (OpenAI) and Claude (Anthropic) and reviewed by the authors.

Table 6. Components of the framework.

#	Component	Meaning
1	Governance architecture	Rules, structures and institutional arrangements
2	Information and understanding	Decision maker understands AI output, limitations and relevant information
3	Decision authority	Decision maker holds formal and practical authority
4	Oversight / intervention / override	Capacity to monitor, intervene and override in AI-supported decisions
5	Human–AI interaction	The actual decision interaction between human and AI
6	Contestability	Possibility to review, question and challenge a decision
7	Organizational enablement	Training, resources, role clarity, monitoring and implementation support
8	Decision outcomes	Decision-level outcomes linked to the governance process

5.2 Meaningful human control

Meaningful human control does not mean the mere presence of a human. It means that the individual can understand, evaluate, challenge, intervene in and override the AI-assisted decision, which in turn requires competence, time and accountability support. Actual decision authority is therefore the hinge on which the framework turns.

5.3 Testable propositions

- **P1.** Stronger governance architecture improves the quality of information available to decision makers and their understanding of AI outputs.
- **P2.** Better information and understanding strengthen decision makers' effective decision authority.
- **P3.** Genuine decision authority, combined with a workable override, increases meaningful human control.
- **P4.** Calibrated trust in AI recommendations improves decision outcomes.
- **P5.** Heavier workload weakens human oversight.
- **P6.** Accessible contestability mechanisms increase the correction of erroneous decisions.
- **P7.** Organizational enablement strengthens the implementation of governance mechanisms.

These propositions require testing through experiments, longitudinal studies and validated measures (Section 6.4).

5.4 Positioning against existing frameworks

The framework builds on, rather than replaces, existing accounts (Table 7). Santoni de Sio and van den Hoven (2018) define meaningful human control through a tracking condition, under which the system responds to the relevant moral reasons of those who design and deploy it, and a tracing condition, under which every outcome can be traced to at least one human in the chain of design and use. Theirs is a philosophical account developed for autonomous systems. We carry the same concern into AI-assisted decisions and break it into components that can be observed and measured in organizations: information and understanding, decision authority, override capacity, competence and timing, and accountability.

The EU Artificial Intelligence Act gives human oversight a legal form. Article 14 requires high-risk systems to be designed so that natural persons can oversee them effectively: they must understand the system's capacities and limitations, remain aware of automation bias, interpret outputs correctly, decide not to use the system or to disregard, override or reverse its output, and interrupt it through a stop procedure (European Parliament & Council of the European Union, 2024). These requirements map closely onto our information, authority and override components. Yet Article 14 specifies what oversight must make *possible*; it does not show whether these capabilities are *exercised* in practice or whether they change outcomes. That is precisely the gap this evidence base exposes.

The voluntary NIST AI Risk Management Framework organizes risk management into four functions (Govern, Map, Measure and Manage) and asks organizations to define roles and processes for human oversight (National Institute of Standards and Technology, 2023). Its focus on organizational risk processes corresponds to our governance architecture and organizational enablement components; our framework adds a decision-level account of what those processes must achieve for the person making the decision. Finally, Papagiannidis et al. (2025) reviewed 48 articles and proposed structural, procedural and relational governance practices linked to antecedents and to effects such as business value, employee well-being and reputation. Their framework explains how organizations govern AI in general. Ours differs in scope and focus: it addresses AI-assisted decision-making specifically, grades 186 studies by level of evidence, and positions meaningful human control as the mechanism between governance arrangements and decision outcomes.

Table 7. Positioning of the present framework.

Account	Level and scope	What it specifies	What the present work adds
---------	-----------------	-------------------	----------------------------

Account	Level and scope	What it specifies	What the present work adds
Santoni de Sio & van den Hoven (2018)	Philosophical; autonomous systems	Tracking and tracing conditions for meaningful human control	Observable, measurable components for AI-assisted organizational decisions
EU AI Act, Art. 14 (2024)	Legal; high-risk AI systems	Capabilities that oversight must enable (understand, disregard, override, stop)	Tests whether capabilities are exercised and change outcomes; evidence grading
NIST AI RMF 1.0 (2023)	Voluntary; organizational risk processes	Govern, Map, Measure and Manage functions	Decision-level account of what processes must achieve for the decision maker
Papagiannidis et al. (2025)	Review of 48 articles; organizational AI governance	Structural, procedural and relational practices; antecedents and effects	Decision-making focus; 186 studies graded by evidence level; control as mediating mechanism

6. Discussion

6.1 Principal findings

The central finding is a gap between how much the field relies on human oversight and how little it knows about whether oversight works. Accountability and transparency are treated almost universally and human-centered control appears in 70% of studies, yet only three studies can support causal inference, and none of the causal studies tests whether oversight improves the quality of decisions. Treating governance as a compliance architecture is therefore insufficient: what matters is the information, authority and capacity for accountable control that governance actually gives to the people making decisions.

6.2 Theoretical contribution

For the literature on meaningful human control, the review offers an empirical counterpart to largely normative accounts (Santoni de Sio & van den Hoven, 2018; Cornelissen et al., 2022). Governance appears in the corpus as a network of interconnected mechanisms that link institutional arrangements to human decision capabilities and outcomes (Hunt & McKelvey, 2019; Koniakou, 2021; Gordon et al., 2022; Hickok, 2022; Lazcoz & De Hert, 2023; Lahusen et al., 2024; Papagiannidis et al., 2025). The central theoretical question is which of these mechanisms shape human decision processes, under what conditions and in what contexts, and whether the control they confer is exercised in practice.

6.3 Methodological and practical contributions

Methodologically, future studies should specify the governance mechanism, the human process, the outcome of interest and the contextual moderators, and should report causal and associational evidence separately. Pre-coding evidence strength by study design would also make meta-analysis and stronger synthesis feasible (Page et al., 2021a, 2021b). *Practically*, the existence of a policy or framework is not proof of success (Watkins et al., 2021; Green, 2022; Alfrink et al., 2023). Training, resources, authority, monitoring, escalation and contestability must actually be implemented (Fan, 2025; Lin et al., 2025; Nousi et al., 2025; Workum et al., 2025), and without organizational readiness and role clarity, formal governance is little more than a paper exercise (Erdemir & Sumbas, 2026; Kim & Shin, 2026; Tekin & Stoycheva, 2026; Widlak, 2026). For public administrations, the implication is that compliance with oversight mandates such as Article 14 of the EU AI Act should be judged by whether officials actually exercise informed authority over AI-assisted decisions, not by whether a human reviewer is formally assigned (Green, 2022).

6.4 Research agenda

Table 8 translates the identified gaps into design-specific priorities (Koulu, 2020; Lyons et al., 2021; Cornelissen et al., 2022; Costanza-Chock et al., 2022; Green, 2022; Waldman & Martin, 2022; Pi, 2024; Papagiannidis et al., 2025; Torrecilla-Pinero, 2026).

Table 8. Prioritized research agenda linked to framework propositions.

Gap	Recommended design	Outcome focus	Proposition
Measuring meaningful human control	Scale development and validation; invariance testing across sectors	Construct validity	P1–P3
Override effectiveness	Randomized override experiments with realistic decision tasks	Error correction, decision quality	P3
Calibrated reliance	Experiments varying AI accuracy and uncertainty displays	Appropriate reliance, automation bias	P4
Workload and proportionality	Field and experimental studies of oversight load by risk tier	Oversight quality, burden	P5
Contestability	Audit studies and field trials of appeal mechanisms	Correction of erroneous decisions	P6
Implementation fidelity	Longitudinal, multi-site implementation studies	Fidelity, sustained practice	P7
Shared accountability	Multi-actor case and network studies	Answerability across actors	P1, P3

6.5 Limitations

The 186 studies are heterogeneous in design, context, constructs and measurement, which limits like-for-like comparison, and not all reported measurement items, reliability and validity data or source locators; where information was missing, we made no inferences. Several limitations of the review process also warrant caution. Title and abstract screening was AI-assisted and then checked by human reviewers; only AI–human agreement was quantified ($\kappa = 0.761$), and no inter-rater statistic for independent human reviewers is reported. The review was not prospectively registered. Of the 406 records sought, 153 (38%) could not be retrieved and so could not be assessed for eligibility, which may have introduced retrieval bias. The search covered only two databases and English-language publications, and it excluded grey literature and policy documents, although relevant evidence may reside in regulations, standards and organizational materials. Half of the corpus could not be rated on conventional risk-of-bias criteria. Finally, causal evidence was very limited (two causal studies and one causal-capable study); the framework is therefore evidence-informed and integrative rather than a validated causal model.

7. Conclusion

Human oversight is now the default safeguard for AI-assisted decisions, but the evidence that it works is thin. The conceptualization and measurement of AI governance remain fragmented, and the evidence linking governance to outcomes is overwhelmingly non-causal: two studies (1.1%) were causal and one (0.5%) causal-capable, and both causal studies measured perceived legitimacy rather than decision quality. The human-centered framework turns this gap into a testable question: to what extent do governance mechanisms give people meaningful opportunities to exercise informed and accountable control over AI-assisted decisions? In answer to the research questions:

- **RQ1.** Governance consolidates accountability, transparency, human oversight, fairness, risk management, regulation, privacy, participation, contestability and organizational mechanisms.

- **RQ2.** Surveys, frameworks and checklists, indices and scales, technical metrics and qualitative approaches have all been used; measurement is fragmented and cross-study comparison remains limited.
- **RQ3.** Positive, negative, mixed, descriptive and conceptual relationships are reported, but causal evidence is very limited (3 of 186 studies; 1.6%).
- **RQ4.** The principal gaps are the measurement of meaningful human control, causal testing of mechanisms, override effectiveness, calibrated reliance, workload, contestability, implementation fidelity and multi-actor accountability.

Until studies of this kind exist, a requirement for human oversight is best understood as a hypothesis about better decisions, not a guarantee of them (Koulu, 2020; Cornelissen et al., 2022; Green, 2022; Lazcoz & De Hert, 2023; Torrecilla-Pinero, 2026).

Data availability

The search strategies (Table S1), the evidence-traceability table (Table S2) and the full list of the 186 included studies (Table S3) are provided in the Supplementary Material. The eligibility criteria and coding framework, the title and abstract screening log, the full-text eligibility decisions, the data extraction matrix, the study-level quality appraisal and the duplicate and scope adjudication record (Tables S4–S9) are provided in a separate Supplementary Data file.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

The research design, literature search, data extraction, quality appraisal, analysis, interpretation of findings and development of the framework were carried out by the authors. Study screening and selection were conducted by the authors with the help of an AI tool for preliminary triage of titles and abstracts; all final screening decisions were made through human review, as described in Section 3.3. During the preparation of the manuscript, the authors used ChatGPT (OpenAI) and Claude (Anthropic) to draft, restructure and revise text on the basis of the authors' analysis and earlier drafts; to edit language and grammar; to improve the organization and presentation of text and tables; to generate the code used to produce the figures from the authors' coded study data; to check the internal consistency of numbers and citations; to verify reference details; and to obtain critical feedback on the manuscript. All AI-assisted content was reviewed, revised and verified by the authors against the underlying sources and study data, and the authors take full responsibility for the content of the published article. No generative AI tool was used to create or alter research data, results or references.

References

- Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (2023). Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, ACM, 16 pages. <https://doi.org/10.1145/3544548.3580984>
- Ali, S., Mohiuddin, K., Sachdeva, S., Juyal, A., Praveen, S., & Khan, A. (2026). Adaptive audit frameworks for autonomous and AI-assisted decision ecosystems. *EDPACS: The EDP Audit, Control, and Security Newsletter*. <https://doi.org/10.1080/07366981.2026.2660882>
- Clavijo, G. M. C., Nong, P., Tan, S., & Platt, J. (2026). Public Expectations for Food and Drug Administration Approval of AI-Based Clinical Decision Support Tools: Quantitative Study. *JMIR AI*, 5, e84315. <https://doi.org/10.2196/84315>
- Cornelissen, N. A. J., van Eerd, R. J. M., Schraffenberger, H. K., & Haselager, W. F. G. (2022). Reflection machines: increasing meaningful human control over Decision Support Systems. *Ethics and Information Technology*. <https://doi.org/10.1007/s10676-022-09645-y>

- Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. <https://doi.org/10.1145/3531146.3533213>
- Devereaux, A. (2025). Hot nudges on hazy landscapes. *Mind & Society*. <https://doi.org/10.1007/s11299-025-00358-5>
- Erdemir, N., & Sumbas, E. (2026). Integrating Artificial Intelligence into Psychological Counseling: A Narrative Review and Governance Framework. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 63, 1--16. <https://doi.org/10.1177/0046958026143832>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Fan, Y. (2025). The role of artificial intelligence in the digital transformation of government: opportunities and ethical challenges. *Frontiers in Public Health*, 13, 1694996. <https://doi.org/10.3389/fpubh.2025.1694996>
- Gordon, G., Rieder, B., & Sileno, G. (2022). On mapping values in AI Governance. *Computer Law & Security Review*, 46, 105712. <https://doi.org/10.1016/j.clsr.2022.105712>
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681. <https://doi.org/10.1016/j.clsr.2022.105681>
- Hauer, M. P., Krafft, T. D., & Zweig, K. (2023). Overview of transparency and inspectability mechanisms to achieve accountability of artificial intelligence systems. *Data & Policy*, 5, e36, Cambridge University Press. <https://doi.org/10.1017/dap.2023.30>
- Hesselmann, F. (2024). Scaffolding decision spaces in decision support systems: Using plagiarism screening software in editorial offices. *Big Data & Society*. <https://doi.org/10.1177/20539517241306371>
- Hickok, M. (2022). Public procurement of artificial intelligence systems: new risks and future proofing. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-022-01572-2>
- Hillo, J., Vento, I., & Erkkilä, T. (2025). Algorithmic Governance: Experimental Evidence on Citizens' and Public Administrators' Legitimacy Perceptions of Automated Decision-Making. *Public Administration*. <https://doi.org/10.1111/padm.70028>
- Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2019). Improving the content validity of the mixed methods appraisal tool: A modified e-Delphi study. *Journal of Clinical Epidemiology*, 111, 49--59.e1. <https://doi.org/10.1016/j.jclinepi.2019.03.008>
- Hulstijn, J. (2023). Computational Accountability. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAAIL 2023)*. <https://doi.org/10.1145/3594536.359512>
- Hunt, R., & McKelvey, F. (2019). Algorithmic Regulation in Media and Cultural Policy: A Framework to Evaluate Barriers to Accountability. *Journal of Information Policy*. <https://doi.org/10.5325/jinfopoli.9.2019.0307>
- Johari, B. (2026). Dual erosion of equality and remedy in algorithmic decision systems. *Discover Artificial Intelligence*, 6, 245. <https://doi.org/10.1007/s44163-026-01040-6>
- Kim, J., & Shin, H. (2026). Stage-Aware Governance of Large Language Models: Managing Uncertainty and Human Oversight in AI-Assisted Literature Review Systems. *Systems*. <https://doi.org/10.3390/systems14020153>
- Koniakou, V. (2021). Governing Artificial Intelligence and Algorithmic Decision Making: Human Rights and Beyond. *Conference proceedings: 20th Conference on e-Business, e-Services and e-Society (I3E)*, Springer/IFIP. https://doi.org/10.1007/978-3-030-85447-8_16
- Koulu, R. (2020). Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law*. <https://doi.org/10.1177/1023263X20978649>
- Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. *Scientific Reports*. <https://doi.org/10.1038/s41598-024-71761-0>
- Lazcoz, G., & De Hert, P. (2023). Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities. *Computer Law & Security Review*. <https://doi.org/10.1016/j.clsr.2023.105833>
- Lin, A. L., Parrish, A. B., Cary, M., Silcox, C., Balu, S., Jelovsek, J. E., O'Brien, C., Pencina, M., Poon, E., & Economou-Zavlanos, N. J. (2025). Algorithm-Based Clinical Decision Support: Evolving Regulatory

- Landscape and Best Practices for Local Oversight. *Annual Review of Biomedical Data Science*.
<https://doi.org/10.1146/annurev-biodatasci-103123-094601>
- Liu, N. (2025). Research on the Guarantee Mechanism of Transparency Rights in Artificial Intelligence Decision-making Oriented to Algorithm Interpretability: A Comparison of Technical Governance Based on GDPR and Beijing Consensus. *Proceedings of the 2025 International Conference on Information Economy, Data Modeling and Cloud Computing (ICIDC 2025)*, ACM. <https://doi.org/10.1145/3772900.3772949>
- Lukkien, D. R. M., Stolwijk, N. E., Askari, S. I., Hofstede, B. M., Nap, H. H., Boon, W. P. C., Peine, A., Moors, E. H. M., & Minkman, M. M. N. (2024). AI-Assisted Decision-Making in Long-Term Care: Qualitative Study on Prerequisites for Responsible Innovation. *JMIR Nursing*. <https://doi.org/10.2196/55962>
- Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, Vol. 5, CSCW1, Article 106. <https://doi.org/10.1145/3449180>
- Natali, C., & Cabitza, F. (2025). Make Some Noise for Ground Truthing! Frictional Design against Epistemic Sclerosis in Decision Support Systems. *Scandinavian Journal of Information Systems*, 37(2), 361--392. <https://doi.org/10.17705/3SJIS/037.22>
- National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- Nouis, S. C., Uren, V., & Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision-making: a qualitative study of healthcare professionals' perspectives in the UK. *BMC Medical Ethics*. <https://doi.org/10.1186/s12910-025-01243-z>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021a). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . McKenzie, J. E. (2021b). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372, n160. <https://doi.org/10.1136/bmj.n160>
- Pan, H., Liu, J., & Liu, S. (2026). Enhancing Physician Resilience to Generative AI: Multilevel Framework for Shared Authority, Verification, and Skill Preservation. *Journal of Medical Internet Research*, 28, e88058. <https://doi.org/10.2196/88058>
- Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *Journal of Strategic Information Systems*. <https://doi.org/10.1016/j.jsis.2024.101885>
- Pi, Y. (2024). Empowering Individuals in Automated Decision-Making: Explainability, Contestability and Beyond. *CSCW Companion '24: Companion of the 2024 Computer-Supported Cooperative Work and Social Computing*. <https://doi.org/10.1145/3678884.36820>
- 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>
- Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., . . . Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, 355, i4919. <https://doi.org/10.1136/bmj.i4919>
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., . . . Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, i4898. <https://doi.org/10.1136/bmj.i4898>
- Tekin, E., & Stoycheva, B. (2026). Algorithmic Management and the Social Sustainability of Employment Relations: Representationless Governance in Platform Courier Labor. *Sustainability*. <https://doi.org/10.3390/su18105011>
- Torrecilla-Pinero, J. A. (2026). Judgment Cannot Be Delegated: A Subject-Preserving Framework for AI Governance. *Philosophy & Technology*, 39, 135. <https://doi.org/10.1007/s13347-026-01140-2>
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*, Paper 440, 1--14. <https://doi.org/10.1145/3173574.3174014>

- Vigeant, T., Tam, A., Shi, Q., Allison, J., Kim, M., Sim, J. A., Won, D. O., Shin, D., McManus, D. M., Lee, J. J., Yu, J. Y., & Zai, A. H. (2026). A Structured Comparison of the Coalition for Health AI Responsible AI Guide and South Korea's Trustworthy AI Guideline for Health Care AI Assurance: Comparative Framework Analysis. *JMIR AI*, 5, e86220. <https://doi.org/10.2196/86220>
- Waldman, A., & Martin, K. (2022). Governing algorithmic decisions: The role of decision importance and governance on perceived legitimacy of algorithmic decisions. *Big Data & Society*. <https://doi.org/10.1177/2053951722110044>
- Watkins, E. A., Moss, E., Metcalf, J., Singh, R., & Elish, M. C. (2021). Governing Algorithmic Systems with Impact Assessments: Six Observations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. <https://doi.org/10.1145/3461702.3462580>
- Widłak, T. (2026). Techno-Legal Virtues in Hybrid Adjudication: Judicial Excellence in AI-Augmented Courts. *Minds and Machines*. <https://doi.org/10.1007/s11023-025-09756-7>
- Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. *Data & Policy*. <https://doi.org/10.1017/dap.2022.39>
- Wilson, C., & van der Velden, M. (2022). Sustainable AI: An integrated model to guide public sector decision-making. *Technology in Society*. <https://doi.org/10.1016/j.techsoc.2022.101926>
- Workum, J. D., van de Sande, D., Gommers, D., & van Genderen, M. E. (2025). Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare. *Frontiers in Artificial Intelligence*. <https://doi.org/10.3389/frai.2025.1504805>
- Yousefi, Y. (2024). Approaching Fairness in Digital Governance: A Multi-Dimensional, Multi-Layered Framework. *Proceedings of the 17th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2024)*, ACM, 7 pages. <https://doi.org/10.1145/3680127.3680197>
- Zieliński, P., Roesti, M., Barreto, R., & Tahaei, M. (2026). Operationalizing Responsible AI Policies With LLMs: An End-To-End Monitoring Prototype. *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26)*, ACM, 15 pages. <https://doi.org/10.1145/3772363.3799675>
- Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 38(3), 101577. <https://doi.org/10.1016/j.giq.2021.101577>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryDataTablesS4S9.xlsx](#)
- [SupplementaryMaterialTablesS1S3.docx](#)