AI & SOCIETY
https://doi.org/10.1007/s00146-026-03349-3

OPEN FORUM

Decision substitution and the reorganisation of judgement in AI-mediated society

Ayodeji Olusegun Ibitoye[1] · Olufunke Faderera Ajenifuja[1] · Razia Sulthana Abdul Kareem[1] · Deborah D. Adeyemo[2] · Rantiola Fidelis Famutimi[3]

Received: 31 March 2026 / Accepted: 16 August 2026
© The Author(s) 2026

Abstract
Artificial intelligence (AI) systems are increasingly embedded in institutional and everyday contexts, where they are publicly framed as decision-support tools yet often function as de facto decision-makers. This study argues that the shift from algorithmic support to decision substitution constitutes a significant societal problem insufficiently addressed by prevailing AI ethics and governance frameworks. Rather than focussing solely on technical performance, bias, or transparency, the analysis examines how AI-mediated practices reconfigure human agency, responsibility, and judgement. Drawing on philosophy of technology and socio-technical scholarship, it conceptualises AI systems as cultural and epistemic infrastructures that shape how knowledge is produced, authority distributed, and action legitimised. It argues that AI increasingly reorganises the decision field by restructuring visibility, validation, participation, and legitimacy, thereby transforming the conditions under which judgement is exercised. The research distinguishes between decision support, decision delegation, and decision substitution, showing how human judgement can be displaced even when human actors remain formally involved. The analysis identifies threshold conditions through which decision substitution emerges, including constrained dissent, diminished corrective capacity, default algorithmic legitimacy, and the separation of authority from accountability. These dynamics generate responsibility asymmetries in which accountability remains attached to human actors while authority migrates towards algorithmic systems. The problem is most acute in contexts characterised by indeterminacy, where values, tacit knowledge, and ethical concerns resist formalisation. The research concludes by advancing a society-shaping-the-algorithm perspective that reframes human-in-the-loop governance as a collective and institutional condition rather than a procedural safeguard, providing a conceptual foundation for preserving meaningful human judgement in AI-mediated societies.

Keywords Decision substitution · AI-mediated judgement · Infrastructural governance · Responsibility asymmetries · Indeterminacy

✉ Ayodeji Olusegun Ibitoye
a.o.ibitoye@greenwich.ac.uk

Olufunke Faderera Ajenifuja
oa50321@gre.ac.uk

Razia Sulthana Abdul Kareem
razia.sulthana@greenwich.ac.uk

Deborah D. Adeyemo
dd.adeyemo@ui.edu.ng

Rantiola Fidelis Famutimi
rfamutimi@unimed.edu.ng

[1] School of Computing and Mathematical Sciences, University of Greenwich, London, UK

[2] Faculty of Law, University of Ibadan, Ibadan, Nigeria

[3] Department of Mathematics and Computer Science, University of Medical Sciences, Ondo, Nigeria

1 Introduction

Artificial intelligence (AI) systems are now deeply embedded in the infrastructures of everyday life: across education, healthcare (Ibitoye et al. 2025a), public administration, and organisational decision-making. AI increasingly mediates how problems are framed, options are prioritised, and actions are legitimised (Kitchin 2019). Increasingly, however, AI systems do more than mediate individual decisions. As they become embedded within organisational routines, professional practices, and everyday interactions, they begin to shape the broader conditions under which decisions are made. Their influence extends beyond recommendations or predictions to encompass what becomes visible, which

Published online: 03 September 2026

forms of reasoning are considered legitimate, and how authority is distributed within decision-making processes.

Public discourse and policy frameworks commonly describe these systems as decision-support technologies, designed to assist human judgement by providing recommendations, predictions, or classifications. Yet, a growing body of empirical research suggests that, in practice, many AI systems function as de facto decision-makers, shaping outcomes in ways that leave limited scope for meaningful human deliberation (Parasuraman and Riley 1997; Eubanks 2018). This study argues that a gradual but consequential transformation is underway. AI systems are increasingly reorganising the decision field itself, restructuring the positions, functions, and sources of legitimacy through which judgements are produced. Decision substitution is not simply the replacement of human decisions by machine outputs; it is the outcome of a broader reconfiguration of the conditions under which judgement becomes possible. This shift is rarely explicit, formally acknowledged, or democratically governed. Instead, it emerges through routinised use, institutional pressures, time constraints, and the perceived authority of algorithmic outputs (Bijker and Law 1994). Human agency is not removed outright, but subtly reconfigured, redistributed, and, in some cases, eroded.

Prevailing AI ethics and governance frameworks struggle to account for this transformation. Much of the literature focuses on algorithmic bias, transparency, fairness, or robustness (Mittelstadt et al. 2016). While essential, these approaches often assume a stable human decision-maker who retains final authority. Less attention has been paid to how AI-mediated practices reshape who decides, how responsibility is assigned, and the conditions under which human judgment remains meaningful and legitimate (Danaher 2022; Rahwan et al. 2019). In particular, existing frameworks rarely examine how algorithmic systems reshape the structural, functional, and normative dimensions of decision-making simultaneously. Questions concerning who occupies authoritative positions, what counts as valid knowledge, and how legitimacy is produced often remain secondary to concerns about technical performance or procedural compliance. The resulting ethical risks extend beyond erroneous or biased outputs to include the normalisation of deference to probabilistic systems whose optimisation goals may diverge from social values. Importantly, these dynamics are most visible not in speculative future scenarios, but in mundane, everyday settings. Empirical studies document professional over-reliance on algorithmic recommendations, institutional embedding of decision systems into workflows, and user deference to ranked or generated outputs perceived as authoritative (Goddard et al. 2012). These practices create responsibility asymmetries in which authority increasingly migrates towards algorithmic systems while accountability remains attached to human actors. Decisions may be justified through reference to system outputs even when those held responsible possess limited capacity to meaningfully influence the conditions under which such outputs are produced.

Against this backdrop, this study examines the societal implications of decision substitution in AI-mediated life. Rather than treating AI as a neutral instrument or an autonomous moral agent, we analyse AI systems as cultural and epistemic infrastructures that reshape agency, responsibility, and the conditions of judgement, aligning with longstanding concerns with algorithmic governance, indeterminacy, and the need to re-centre human values in technological futures (Coeckelbergh 2020). Central to the analysis is the concept of the decision field, understood as the socio-technical environment within which authority, participation, validation, and legitimacy are organised. By examining how AI systems reorganise this field, the study moves beyond questions of automation to explore how infrastructures shape the practical exercise of judgement itself. Its objective is not to present new primary empirical data, but to develop a theoretically integrated account of how AI infrastructures reshape decision-making practices and generate conditions under which decision substitution emerges across institutional and everyday contexts. The analysis foregrounds values, lived experience, and institutional context rather than technical optimisation or system performance.

The argument integrates five complementary bodies of scholarship. First, philosophy of technology provides conceptual tools for analysing agency, autonomy, and moral responsibility in technologically mediated action (Verbeek 2024). Second, sociology and Science and Technology Studies (STS) frame AI systems as infrastructures and institutions that organise social practices, distribute authority, and stabilise norms (Bowker and Star 1999). Third, human–computer interaction (HCI) research contributes empirical insights into automation bias, user reliance, and the dynamics of human–AI interaction (Norman 2013). Fourth, legal and governance scholarship addresses accountability gaps and the limits of existing regulatory models when decision-making is distributed across humans and machines (Danaher 2022; Floridi et al. 2018). Moreover, data science and AI systems' literature informs the discussion of optimisation logics, decision pipelines, and the framing of AI as "support" within organisational discourse (Rahwan et al. 2019). This interdisciplinary synthesis enables the research to examine AI-mediated decision-making as a socio-technical phenomenon rather than a purely technical or moral problem.

Although conceptual in orientation, the study is grounded in existing empirical research on everyday AI use. Throughout the analysis, we draw on qualitative and mixed-methods studies of AI deployment in professional, organisational, and public-sector contexts, including research on algorithmic management, welfare systems, and automated assessment

tools (Eubanks 2018). These studies demonstrate that the shift towards decision substitution is not hypothetical, but embedded in routine practices shaped by institutional constraints and power relations. These suggest that decision substitution rarely occurs through the complete removal of human actors from decision-making processes. Instead, it emerges through more subtle transformations in visibility, validation, participation, and legitimacy that progressively reorient judgement around algorithmic outputs. Hence, understanding these transformations requires attention not only to decisions themselves but also to the infrastructures and institutional arrangements that make particular forms of judgement possible.

This secondary empirical grounding ensures that the argument remains attentive to lived experience, how AI systems are encountered, trusted, and deferred to, rather than abstracted notions of risk or control. Thus, human-in-the-loop is not conceptualised as a technical safeguard but as a societal and normative challenge. Prior research shows that the mere presence of a human actor does not guarantee meaningful oversight, particularly when organisational incentives, workflow constraints, and epistemic authority favour algorithmic outputs (Seaver 2019). The analysis reframes human-in-the-loop governance as a question of participation, accountability, institutional design, and the preservation of meaningful human judgement under conditions of increasing algorithmic mediation.

The research makes four contributions. First, it conceptualises AI systems as cultural and epistemic infrastructures that shape the conditions of judgement rather than merely supporting decisions. Second, it introduces the concept of decision field reorganisation to explain how AI restructures positions, functions, and sources of legitimacy within decision-making processes. Third, it develops a theory of decision substitution and identifies threshold conditions under which human judgement becomes progressively displaced. Then, it advances a responsibility asymmetry perspective that explains why accountability remains attached to human actors even as authority increasingly migrates towards algorithmic systems.

2 Conceptual background: mediation, delegation, and responsibility

Philosophy of technology has long challenged the view that technologies are neutral instruments serving pre-given human intentions. Accounts of technological mediation emphasise that tools shape perception, interpretation, and action, thereby participating in the constitution of agency itself (Ihde 1990; Verbeek 2024). From this perspective, the ethical significance of AI systems does not lie solely in their technical capacities, but in how they reorganise the conditions under which judgement is exercised. Agency is neither eliminated nor replaced by technology; it is redistributed through socio-technical arrangements.

Science and Technology Studies (STS) extend this insight by showing how technologies become infrastructural over time. Once embedded in institutional routines, classification systems and decision tools recede into the background of practice while simultaneously structuring what counts as knowledge, competence, and legitimate action (Bowker and Star 1999). AI systems increasingly operate in this infrastructural mode. Through ranking, scoring, recommendation, and generation, they shape decision environments rather than merely contribute to them. As such systems become normalised, alignment with their outputs may come to define responsible or defensible conduct.

Research on automation bias and human–AI interaction further complicates the assumption that human oversight guarantees meaningful control. Empirical studies show that individuals often defer to algorithmic outputs, particularly under conditions of cognitive load or institutional constraint (Parasuraman and Riley 1997; Gaube et al. 2021). Human roles may shift from deliberation to validation, even where formal authority remains intact. These dynamics suggest that the ethical question is not only whether AI systems are accurate or transparent, but also how they reconfigure the practice of judgement itself.

Responsibility scholarship has described emerging "responsibility gaps" in socio-technical systems, where outcomes cannot be easily attributed to any single actor (Danaher 2022; Ibitoye et al. 2025b). Yet, such gaps are often treated as problems of attribution or oversight failure. The present analysis shifts the focus towards infrastructural embedding. When decision-making becomes structured around algorithmic outputs, responsibility may persist formally while substantive deliberative agency contracts. The resulting tension between accountability and control signals a deeper transformation in the architecture of judgement. Against this background, the following sections examine how AI systems move from mediating decisions to substituting for judgement, and why this shift has distinct normative consequences. Rather than treating substitution as a matter of automation level alone, the analysis locates it in the reorganisation of epistemic authority and institutional practice.

3 AI systems as infrastructures of judgement

AI systems increasingly function not merely as tools within decision-making processes but as epistemic infrastructures that structure how judgement is exercised and recognised. As these systems become embedded in institutional routines, they do not simply assist deliberation; they reorganise it.

These developments raise important questions about how judgement is organised, exercised, and legitimised within AI-mediated environments. Understanding these questions requires examining both the infrastructural role of AI systems and their effects on everyday experiences of agency and decision-making.

3.1 AI systems as cultural and epistemic infrastructures

To understand the societal implications of decision substitution, it is necessary to move beyond treating AI systems as discrete technical artefacts and instead conceptualise them as cultural and epistemic infrastructures. Infrastructures do not merely support action; they actively shape what forms of action, knowledge, and justification become possible and legitimate within particular social contexts (Boddu 2025). When AI systems are embedded into institutional routines and everyday practices, they reorganise how knowledge is produced, how authority is distributed, and how decisions are rendered intelligible. Fundamentally, infrastructures do not merely shape decisions after they occur; they shape the conditions under which decisions become possible in the first place. They influence what can be seen, what can be known, which forms of expertise are recognised, and how legitimacy is constructed within institutional settings. AI systems should be understood not simply as tools acting within decision-making processes, but as socio-technical arrangements that increasingly structure the environment within which judgement is exercised.

From a socio-technical perspective, contemporary AI systems increasingly function as epistemic authorities. Through classification, ranking, recommendation, and generation, they produce outputs that are routinely treated as credible inputs into decision-making, often without transparent articulation of underlying assumptions, limitations, or normative commitments (Deroncele-Acosta et al. 2024). In doing so, AI systems reshape not only decision outcomes, but also the conditions under which something counts as evidence, expertise, or reasonable judgement. This dynamic is particularly pronounced where algorithmic outputs are framed as objective or data-driven, obscuring the historically situated and value-laden choices embedded in system design, data selection, and optimisation goals (Kitchin 2019). This shift is significant because judgement is never exercised in isolation. As philosophy of technology has long argued, technologies mediate perception, interpretation, and action rather than merely assisting them. AI systems matter not only because they generate recommendations, but also because they increasingly shape the horizons within which recommendations acquire meaning. Their influence operates before deliberation by organising the informational, temporal, and institutional conditions under which judgement occurs.

The infrastructural character of AI becomes especially visible through its institutional embedding. Once integrated into organisational workflows, such as risk assessment, automated screening, or generative decision aids, AI systems recalibrate expectations of speed, consistency, and efficiency. Human judgement is increasingly exercised within parameters set by algorithmic systems rather than independently of them (Christin 2017). Over time, this can redefine professional competence, such that alignment with algorithmic outputs becomes a marker of responsible or defensible practice, while deviation demands justification and carries reputational or organisational risk. From the perspective of philosophy of technology, this process reflects broader patterns of technological mediation through which infrastructures shape the practical exercise of judgement and establish expectations regarding what constitutes responsible or legitimate action (Chiodo 2023). As AI-mediated judgements become routine, their influence recedes into the background of practice, rendering them less visible and more difficult to contest. What appears as neutral decision-making is often the stabilisation of particular ways of seeing, valuing, and acting.

Beyond questions of institutional embedding, the significance of AI infrastructures also lies in how they reshape the practical exercise of judgement itself. The concept of infrastructural judgement helps clarify this transformation. Infrastructural judgement refers to how decisions emerge within environments pre-structured by algorithmic systems and institutional expectations rather than through isolated acts of individual deliberation. Traditional accounts often treat judgement as the outcome of individual deliberation. Yet, when AI systems become embedded within organisational routines, judgement increasingly emerges within environments already structured by algorithmic classifications, rankings, predictions, and recommendations. In such contexts, AI does not merely participate in judgement; it helps define the conditions under which judgements become intelligible, defensible, and actionable. The significance of AI lies not solely in what it decides, but in how it reorganises the background conditions that make decision-making possible.

The implications of infrastructural judgement extend beyond decision-making processes to broader questions of power, exclusion, and epistemic authority. Crucially, AI's epistemic role is inseparable from questions of power, exclusion, and epistemic hierarchy. Because AI systems operate on what can be formalised, measured, and historically encoded, they privilege certain forms of knowledge while marginalising others. Tacit understanding, contextual sensitivity, and culturally situated values are frequently rendered invisible or treated as noise. As

Solak (2025) argues, algorithmic authority increasingly reshapes epistemic norms within the social sciences, redefining what is recognised as legitimate knowledge. Relatedly, Hosseini and Sakhaei (2025) show how AI-mediated knowledge production can reinforce global hierarchies, reproducing asymmetries between those who define data and models and those who are subjected to them. Recent scholarship has called for more epistemically reflexive responses to AI infrastructures. Popow (2026) emphasises the pedagogical implications of AI-driven transformations, arguing for the decolonisation of knowledge practices shaped by algorithmic systems. Kelly's (2025) concept of situated epistemic infrastructures similarly highlights how AI operates within post-coherence knowledge regimes, where authority is distributed across technical, institutional, and cultural arrangements. Khurana (2025) further advocates integrating indigenous knowledge systems into AI discourse, foregrounding epistemic pluralism and ethical responsibility in the design and governance of AI systems.

This infrastructural framing also helps explain why responsibility becomes diffuse in AI-mediated decision-making. When AI systems are treated as background conditions of action rather than contestable socio-technical arrangements, their influence is naturalised. Decisions come to appear as follows: "what the system says," even though those systems embody specific institutional priorities, value assumptions, and governance choices (Yeung 2018). The result is not the disappearance of human agency, but its redistribution across arrangements that obscure lines of accountability and complicate moral responsibility. By conceptualising AI systems as cultural and epistemic infrastructures, these findings foreground the deeper societal stakes of decision substitution. The concern is not only that AI systems may err or discriminate, but also that they increasingly shape how societies know, decide, and justify action. By conceptualising AI systems as cultural and epistemic infrastructures, the analysis foregrounds the deeper societal stakes of AI-mediated decision-making. The concern is not only that AI systems may produce errors or reinforce bias, but also that they increasingly shape the conditions under which societies know, judge, and justify action. Understanding AI as infrastructure shifts attention from individual decisions to the broader environments within which decisions become possible. These environments do not merely influence decisions after they are made; they structure the positions, functions, and sources of legitimacy through which judgement is exercised. The next section develops this argument by examining how AI systems reorganise the decision field itself and how such reorganisation creates the conditions under which decision substitution emerges.

3.2 Lived experience and the erosion of human agency

The infrastructural character of AI is not only observable at the level of institutions and governance, but also it is experienced in practice by individuals whose judgement is increasingly mediated by algorithmic systems. If AI infrastructures shape the conditions under which decisions become possible, then their effects should also be visible in everyday experiences of reliance, deference, expertise, and responsibility. Examining these lived experiences provides an important empirical grounding for the conceptual argument developed in this work. However, the lived experience of AI-mediated decision-making is not uniform. The influence of algorithmic systems manifests across different settings and forms of interaction, ranging from everyday reliance on recommendations to the reshaping of professional expertise and the normalisation of algorithmic authority.

To illustrate how infrastructural judgement becomes embedded in practice, the following subsections examine four interconnected manifestations of this process: automation bias and everyday deference, professional judgement under algorithmic constraint, epistemic dependence on generative systems, and the gradual normalisation of agency loss. These examples demonstrate how AI infrastructures influence not only what decisions are made, but also how individuals come to understand, exercise, and justify judgement in increasingly AI-mediated environments.

3.2.1 Automation bias and everyday deference

A substantial body of empirical research documents the phenomenon of automation bias, whereby individuals disproportionately trust and rely on automated recommendations, even when contradictory information is available (Romeo and Conti 2025; Kazim and Tomlinson 2023). This bias is not merely a cognitive error; it is shaped by organisational norms, time pressure, and the cultural authority granted to algorithmic systems. In such contexts, deferring to AI outputs is often experienced as the rational or responsible course of action, particularly when deviation requires additional justification or exposes individuals to blame. Automation bias does not imply passivity or ignorance on the part of users. Rather, it reflects rational adaptation to environments in which algorithmic systems are framed as more reliable, objective, or institutionally sanctioned than human judgement. As a result, decision substitution can emerge even when individuals remain technically capable of independent reasoning.

3.2.2 Professional judgement under algorithmic constraint

Empirical studies of AI use in professional settings, such as healthcare, social services, education, and organisational management, illustrate how algorithmic systems reshape the exercise of expertise. In healthcare, for instance, clinical decision-support systems create powerful "nudges" that can deskill clinicians, with algorithmic outputs becoming the anchor around which diagnostic reasoning is reorganised (Cabitza et al. 2017; Lebovitz et al. 2021). Similarly, in social services, risk-assessment tools become the central organising logic for caseworkers, structuring investigations and shifting street-level expertise towards algorithmic interpretation (Eubanks 2018). Professionals often encounter AI recommendations as defaults around which decision-making is organised, rather than as optional inputs. This is evident in organisational management, where algorithmic systems set defaults for tasks and evaluations, redefining managerial expertise as compliance with and interpretation of system outputs (Kellogg et al. 2020).

In education, AI-driven platforms reconfigure teachers' roles, positioning them as managers of algorithmic processes where system alerts become the default framework for pedagogical action (Williamson and Eynon 2020; Perrotta and Selwyn 2020). Over time, professional competence may become tacitly defined as alignment with algorithmic outputs, while deviation is treated as exceptional, risky, or requiring extensive documentation. This reconfiguration of expertise has important implications for agency. While professionals may remain legally accountable for decisions, their capacity to exercise judgement is constrained by system design, institutional expectations, and performance metrics. Responsibility is thus experienced as asymmetric: humans bear the consequences of decisions, but lack full control over the conditions under which those decisions are made.

3.2.3 Users, generative systems, and epistemic dependence

Beyond institutional settings, generative AI systems increasingly mediate everyday practices of writing, learning, searching, and sense-making (Bender et al. 2021; Noble 2018). Their human-like fluency creates a powerful "illusion of coherence" that positions them as active participants in intellectual labour (Birhane et al. 2023). Consequently, users frequently describe these systems as authoritative, fluent, and persuasive, even when aware of their limitations. Empirical studies show that this persuasive quality endures even when users are explicitly sceptical, fundamentally altering how people evaluate information (Jakesch et al. 2023).

This creates a form of epistemic dependence, in which individuals rely on AI outputs to frame problems, generate options, or validate understanding. Such dependence does not necessarily entail blind trust. Instead, it reflects a gradual narrowing of epistemic horizons, where certain forms of inquiry or reflection are displaced by interaction with systems optimised for plausibility and efficiency rather than truth or ethical judgement. In this sense, decision substitution operates not only at the level of outcomes but also at the level of thinking itself.

3.2.4 Normalisation, moral deskilling, and agency loss

As AI-mediated practices become routine, their influence recedes into the background of everyday life. This infrastructural invisibility contributes to what has been described as moral deskilling, whereby opportunities for ethical reflection and judgement diminish as decisions are increasingly pre-structured by systems. Individuals may experience a reduced sense of authorship over decisions, even while remaining formally responsible for their consequences. Crucially, this erosion of agency is unevenly distributed. Marginalised groups are often subject to more intensive forms of algorithmic scrutiny and control, while having fewer opportunities to contest or override system outputs. Lived experience thus reveals that decision substitution is not merely a technical shift, but a socio-political one, with implications for power, voice, and inclusion.

3.2.5 Implications for the conceptual framework

These empirical insights reinforce the value of distinguishing between decision support, delegation, and substitution. They show that substitution is not defined solely by system autonomy, but by how AI-mediated practices are experienced by users and professionals. When deliberation is compressed, dissent discouraged, and responsibility individualised, AI systems effectively substitute for human judgement, even in the absence of full automation. These findings suggest that decision substitution is not best understood as a technological event but as a lived socio-technical process. Individuals continue to participate in decision-making, yet their capacity to shape, contest, and author judgements becomes increasingly conditioned by algorithmic infrastructures. These experiences provide a foundation for the next section, which examines how such transformations can be understood as a broader reorganisation of the decision field.

4 The reorganisation of the decision field

4.1 Structural reorganisation of the decision field

The concept of decision substitution requires a broader analytical lens than the transfer of discrete decision tasks from humans to machines. Such a view risks obscuring a more

fundamental transformation occurring within AI-mediated environments: the reorganisation of the decision field itself. Decisions do not emerge in isolation but are produced through relationships among actors, institutional rules, sources of authority, and mechanisms of validation, a view consistent with socio-technical and institutional perspectives that understand decision-making as embedded within networks of human and non-human actors (Lehto et al. 2021). Understanding the societal implications of AI requires examining how these relationships are reconfigured when algorithmic systems become embedded within decision processes. This perspective aligns with socio-technical and institutional approaches that understand decision-making as embedded within networks of human and non-human actors rather than as the outcome of isolated individual choices (Jasanoff 2004; Latour 2005).

Traditional accounts of decision-making typically assume a relatively stable distribution of roles. Human actors identify relevant information, interpret its significance, weigh competing considerations, and ultimately assume responsibility for outcomes. Technologies may support these activities, but they do not fundamentally alter the structure of authority within which judgement occurs. In AI-mediated contexts, however, this assumption becomes increasingly difficult to sustain. The integration of predictive systems, recommendation engines, ranking algorithms, and generative AI tools introduces new centres of epistemic authority that reshape how decisions are produced and justified (Beer 2019; Couldry and Hepp 2018). A useful way of understanding this transformation is through the changing positions occupied by actors within the decision field. As AI systems become increasingly influential, human actors often shift from the role of judge to that of validator. Rather than originating judgements through independent deliberation, they are increasingly asked to review, confirm, or approve recommendations that have already been generated elsewhere. The formal presence of human oversight remains intact, yet the substantive locus of judgement has begun to migrate. Human intervention frequently occurs after the system has structured available options, prioritised information, and established the parameters within which disagreement can occur.

A second shift occurs from agent to operator. In traditional decision environments, actors exercise discretion through interpretation, negotiation, and contextual reasoning. In AI-mediated settings, however, discretion is increasingly channelled through predefined workflows and algorithmically structured procedures. The role of the human actor becomes less about determining what ought to be done and more about ensuring that organisational processes are completed efficiently and consistently. Decision-making becomes operationalised, with human actors managing system outputs rather than exercising autonomous judgement. Perhaps the most significant transformation concerns the relationship between authority and responsibility. Although algorithmic systems increasingly influence outcomes, responsibility often remains attached to human actors. Individuals continue to be held accountable for decisions that are substantively shaped by systems they neither designed nor fully control. This creates a structural asymmetry in which authority migrates towards algorithmic infrastructures while responsibility remains human, echoing concerns regarding distributed agency and accountability in automated decision systems (Langer 2024). The result is not the disappearance of accountability, but its redistribution across actors whose capacity to influence outcomes no longer corresponds to the obligations they bear.

These shifts reveal that decision substitution is not adequately understood as a binary question of whether a machine or a human makes a decision. Instead, it involves a reorganisation of the positions through which decisions are produced, authorised, and justified. AI systems increasingly function as epistemic intermediaries that shape the conditions under which judgement is exercised, while human actors occupy roles oriented towards validation, implementation, and consequence management. The significance of this transformation is not only in who decides, but also in how the architecture of decision-making itself is reconstructed. This structural perspective also helps explain why the societal consequences of AI often remain difficult to identify. The displacement of judgement rarely appears as an explicit transfer of authority. Rather, it emerges gradually through institutional practices that normalise new distributions of responsibility and expertise. Human actors remain present within the decision process, but their position within that process changes. Therefore, the challenge is not the replacement of humans by machines, but the subtle reconfiguration of the roles through which judgement, authority, and accountability are enacted.

4.2 Functional reorganisation of the decision field

If the structural effects of AI are visible in the redistribution of roles and authority, its functional effects are evident in the reconfiguration of the processes through which decisions are produced. AI systems do not merely provide information for decision-making; they increasingly shape the conditions under which information becomes visible, actionable, and legitimate. Their influence extends beyond the outcomes of decisions to the organisation of decision-making itself. One important mechanism through which this occurs is the restructuring of visibility. Every decision depends upon what can be seen, recognised, and treated as relevant. AI systems increasingly mediate this process by determining which signals are highlighted, which risks are prioritised, and which possibilities remain invisible. This reflects broader concerns that algorithmic systems actively shape attention and

knowledge production rather than merely presenting information neutrally (de la Torre et al. 2026; Amoore 2020). Recommendation systems, predictive analytics, risk-scoring tools, and generative AI systems all operate through forms of selective visibility. They do not simply present reality; they organise it. In doing so, they establish a preliminary interpretation of the decision environment before human deliberation begins. What appears as neutral information support is often a prior structuring of attention that shapes subsequent judgement.

A second mechanism involves the reorganisation of decision timing. Human judgement traditionally unfolds through processes of reflection, consultation, and contextual interpretation. AI systems, by contrast, are frequently introduced to accelerate decision-making, reduce uncertainty, and increase efficiency. Such developments reflect broader organisational trends in which speed and optimisation increasingly become governing principles of institutional practice (Rosa 2013; Zuboff 2023). Here, decisions become increasingly expected to conform to the speed of computational systems, compressing opportunities for deliberation and encouraging reliance on readily available outputs. In many organisational settings, the question is no longer simply what decision should be made, but whether there is sufficient justification to delay acting on the system's recommendation. AI systems also reshape the processes of validation through which decisions are justified and defended. In conventional decision environments, legitimacy often emerges from professional expertise, contextual knowledge, or collective deliberation. In AI-mediated environments, however, algorithmic outputs frequently become reference points against which alternative judgements are evaluated, reinforcing forms of automation bias and epistemic dependence documented across professional settings (Parasuraman and Riley 1997; Goddard et al. 2012).

Decisions are increasingly considered reasonable because they align with system recommendations, while deviations require additional explanation. The burden of justification shifts. Rather than the system needing to demonstrate why its recommendation should be accepted, human actors may find themselves required to explain why it should be rejected. This shift has important implications for participation. Decision-making is not merely a technical process but a social activity involving the capacity to intervene, contest, and influence outcomes. Yet, the growing reliance on algorithmic systems can narrow the space within which meaningful participation occurs, particularly where affected individuals possess limited opportunities to understand, challenge, or influence algorithmic recommendations (Benjamin 2023). Individuals affected by decisions may have limited visibility into how recommendations are generated, while professionals operating within AI-mediated environments may encounter organisational pressures that discourage deviation from system outputs. Participation remains formally possible, but its practical significance may be diminished. The ability to intervene becomes increasingly separated from the ability to alter outcomes.

These functional transformations reveal that AI systems do not simply operate within pre-existing decision environments. They actively reorganise the conditions through which judgement is exercised. By shaping visibility, accelerating decision timing, establishing validation criteria, and redefining participation, AI systems influence not only what decisions are made but also how decisions become thinkable, defensible, and actionable in the first place. This perspective helps explain why decision substitution often occurs gradually and without explicit recognition. Human actors continue to participate in decision processes, yet the functional architecture of those processes has already been transformed. The significance of AI lies not solely in its capacity to generate recommendations, but in its ability to structure the environment within which judgement itself is exercised. Understanding this transformation is essential for explaining why the effects of AI extend beyond questions of accuracy or efficiency to encompass deeper changes in authority, agency, and legitimacy.

4.3 Normative reorganisation of the decision field

The structural and functional transformations associated with AI-mediated decision-making are accompanied by a deeper normative shift: the reorganisation of the sources through which decisions acquire meaning, legitimacy, and authority. This concern resonates with broader scholarship on technological mediation, epistemic authority, and the social production of legitimacy (Fricker 2017). Decision substitution does not occur simply because AI systems are faster, more efficient, or more accurate. Rather, it occurs because algorithmic outputs increasingly become recognised as legitimate grounds for action. The central challenge is not only technological but also normative. It concerns how societies come to trust, justify, and defer to particular forms of judgement. Traditionally, the legitimacy of decisions has been linked to identifiable human capacities, including practical wisdom, contextual interpretation, and professional expertise (Arendt 2019; Ricoeur 1992).

Professional expertise, practical wisdom, contextual understanding, and moral reasoning have provided the basis upon which decisions are considered credible and defensible. While these forms of legitimacy have never been uncontested, they are rooted in the assumption that judgement emerges through processes of interpretation and deliberation undertaken by accountable actors. In AI-mediated environments, however, legitimacy increasingly derives from a different source: the perceived objectivity of computational systems. Algorithmic systems are frequently presented as

data-driven, evidence-based, and impartial. Their authority is often reinforced through narratives of objectivity and neutrality that obscure the social and political assumptions embedded within technical systems (Luusua 2023). Their recommendations often carry an aura of neutrality that exceeds the perceived reliability of individual human judgement. Hence, this legitimacy is not solely a consequence of technical performance. Many users remain unable to evaluate the internal operation of complex AI systems, yet still regard their outputs as credible.

The authority of the system rests not only on what it does, but also on what it symbolically represents. AI becomes associated with consistency, rationality, and freedom from subjective bias, even when such assumptions are themselves contestable. This symbolic dimension is crucial for understanding decision substitution. Human actors do not simply follow algorithmic recommendations because they are compelled to do so. They frequently do so because those recommendations appear more defensible than alternative judgements. In institutional settings, aligning with system outputs may be perceived as prudent, responsible, and professionally legitimate. Conversely, disagreement may expose individuals to greater scrutiny, additional justification requirements, or retrospective accountability if outcomes prove unfavourable. The result is a subtle redistribution of legitimacy in which algorithmic outputs become the default reference point for determining what constitutes a reasonable decision.

At this stage, the role of the human actor undergoes an important transformation. Individuals remain present within decision processes, yet they increasingly cease to function as the primary source of judgement. Their role becomes one of confirmation, implementation, or procedural oversight. Judgement is not formally transferred to the system, but its practical authority shifts towards algorithmically generated recommendations. Human participation persists, while human authorship diminishes. The normative consequences of this transformation extend beyond questions of responsibility. They affect the very meaning of judgement itself. Judgement has traditionally involved the capacity to navigate ambiguity, weigh competing values, and exercise discretion under conditions of uncertainty. It is not merely the application of rules to facts, but an interpretive activity through which individuals and institutions negotiate what ought to be done.

When legitimacy becomes increasingly attached to algorithmic outputs, the scope for such interpretive work may narrow. Decisions become valued less for the quality of deliberation that produced them and more for their alignment with system-generated assessments. This dynamic is particularly significant because legitimacy tends to reproduce itself through institutionalisation and routine practice, becoming embedded within organisational cultures and expectations over time (Berger and Luckmann 2016; Suchman 1995). As algorithmic recommendations are repeatedly accepted, they become institutionalised as common-sense reference points. Their authority no longer depends on continuous justification; it becomes embedded within organisational expectations, professional norms, and routine practices. Over time, the legitimacy of the system is reinforced through its very use. What begins as a recommendation gradually acquires the status of a standard against which alternative judgements are evaluated. From this perspective, decision substitution should not be understood solely as a redistribution of tasks between humans and machines. It is also a transformation in the social production of meaning.

AI systems increasingly participate in defining what counts as a valid judgement, who is recognised as an authoritative decision-maker, and which forms of reasoning are considered legitimate. The displacement of judgement occurs not because humans disappear from decision-making processes, but because the normative foundations of judgement are progressively reorganised around algorithmic forms of authority. This normative reorganisation completes the transformation of the decision field. Structurally, positions are redistributed. Functionally, the conditions of judgement are reorganised. Normatively, legitimacy and meaning migrate towards algorithmic systems. These shifts help explain why decision substitution can emerge even in environments where human actors remain formally responsible and visibly present. The issue is not simply who decides, but how authority, legitimacy, and judgement are reconstructed in AI-mediated societies.

Algorithmic legitimacy becomes stabilised through a recursive institutional process rather than through technical performance alone. Systems are commonly introduced through narratives of efficiency, consistency, evidence-based practice, or risk reduction. Once embedded in workflows, repeated reliance transforms their outputs from optional recommendations into familiar organisational reference points. Alignment with those outputs then becomes easier to document and defend, while deviation attracts additional scrutiny. Policies, performance measures, audit procedures, training practices, and professional norms may subsequently incorporate the system's recommendations as indicators of reasonable conduct. The accumulated history of institutional use is then treated as evidence of reliability and legitimacy, reinforcing further reliance. Through this cycle, algorithmic authority becomes normalised: previous acceptance justifies continued use, continued use stabilises institutional expectations, and those expectations increase the cost of alternative judgement. Legitimacy is reproduced not simply because users believe that AI is objective, but because organisational arrangements repeatedly reward conformity and make system-aligned decisions more administratively and professionally defensible.

4.4 Illustrative vignette: reorganisation of the decision field in healthcare triage

Consider a hospital emergency department that adopts an AI-assisted triage system designed to prioritise patients according to predicted clinical risk. The system analyses patient records, vital signs, demographic information, and historical outcomes to generate a risk score that recommends the urgency of treatment. Formally, clinicians retain responsibility for all decisions, and the AI system is presented as a decision-support tool intended to improve efficiency and consistency. From a structural perspective, however, the introduction of the system alters the positions occupied within the decision process. Before implementation, triage nurses and physicians exercised primary responsibility for assessing urgency, interpreting symptoms, and determining treatment priorities. Following implementation, the system increasingly occupies a role as an epistemic authority. Human clinicians remain present, but their role shifts towards validating, confirming, or occasionally overriding algorithmic recommendations. Responsibility remains attached to medical professionals, while authority becomes increasingly distributed across human and algorithmic actors.

The functional effects are equally significant. The AI system determines which patients appear most urgent, thereby shaping clinical visibility. Certain indicators become prioritised because they are recognised by the model, while other contextual factors may receive less attention. The system also restructures decision timing. Clinicians operating in high-pressure environments may rely on risk scores to accelerate assessments, reducing opportunities for extended deliberation. Validation practices change as well. Decisions that align with the system's recommendation are easily justified as evidence-based, whereas departures from the recommendation require additional explanation and documentation. The most profound transformation occurs at the normative level. Over time, the legitimacy of triage decisions becomes increasingly associated with conformity to algorithmic outputs. Clinicians may continue to exercise judgement, but the perceived defensibility of that judgement is increasingly measured against the system's recommendation. A physician who follows the algorithm and encounters an adverse outcome may be viewed as having acted reasonably because recognised procedures were followed. By contrast, a physician who overrides the recommendation may face greater scrutiny, even when acting on contextual knowledge unavailable to the system. The practical authority of judgement begins to migrate towards the algorithm, despite the formal retention of human oversight.

Thus, no single moment marks the transition from support to substitution. Human actors remain visibly involved throughout the process. Yet, the decision field has been reorganised. The system influences what becomes visible, how decisions are justified, when action occurs, and which forms of reasoning are regarded as legitimate. The result is not the elimination of human judgement, but its gradual displacement within an environment increasingly structured by algorithmic forms of authority. This vignette illustrates the core argument of the study: decision substitution should not be understood as the replacement of humans by machines. Rather, it emerges through the cumulative reorganisation of structural positions, functional processes, and normative expectations that collectively shape how judgement is exercised in practice.

5 Decision substitution and the displacement of judgement

5.1 Beyond assistance: the gradual migration of judgement

The reorganisation of the decision field described in the previous section provides the conditions under which decision substitution emerges. This process is rarely abrupt. Contrary to popular narratives of automation, the displacement of judgement seldom occurs through the complete removal of human actors from decision-making processes. Rather, it unfolds gradually as algorithmic systems become increasingly involved in shaping visibility, validation, legitimacy, and participation. Human actors remain present throughout much of this transition, yet their relationship to judgement changes fundamentally (Brey 2005). This distinction is important because discussions of AI often frame the problem in binary terms: either humans make decisions or machines do. Such formulations obscure the complex sociotechnical arrangements through which authority is redistributed in practice (Latour 2005). In many contemporary settings, AI systems neither fully replace nor merely assist human decision-makers. Instead, they progressively reorganise the conditions under which judgement is exercised, creating environments in which algorithmic outputs become increasingly central to the production and justification of decisions.

The displacement of judgement should be understood as a continuum rather than a discrete event. At one end of this continuum, AI systems function primarily as informational resources that support human deliberation. At the other end, human actors remain formally responsible for outcomes but exercise limited influence over the substantive direction of decisions. Between these poles lies a range of intermediate arrangements in which authority, discretion, and accountability are distributed unevenly across human and algorithmic actors. What distinguishes decision substitution from other forms of human–AI interaction is not the presence of

automation itself, but the migration of epistemic authority (Alkhatib and Bernstein 2019; Green and Chen 2019). As AI systems increasingly define what is relevant, what is reasonable, and what is defensible, human judgement becomes progressively oriented around algorithmic outputs. The result is not necessarily the elimination of human participation but the erosion of human authorship within decision processes. To clarify this progression, the following subsection introduces a typology of decision relations that distinguishes support, delegation, and substitution as analytically distinct configurations within the broader decision field.

5.2 A typology of human–AI decision relations

The distinction between decision support, decision delegation, and decision substitution is not intended as a classification of technologies. The same AI system may occupy different positions across different institutional contexts, depending on how authority, discretion, and accountability are organised. The typology refers to configurations of decision-making rather than technical characteristics of AI systems themselves (Suchman 2007; Orlikowski 2007). Viewed in this way, the significance of AI lies not in what the technology is, but in how judgement is distributed between human and algorithmic actors. The three configurations identified below represent different arrangements of authority and responsibility within the decision field, ranging from situations in which AI informs human judgement to those in which algorithmic outputs become the primary basis upon which decisions are made and justified.

5.2.1 Decision support

In decision-support arrangements, AI systems function primarily as informational resources. They provide recommendations, predictions, or analytical insights that inform human deliberation, but the authority to interpret, evaluate, and act upon those outputs remains substantially with human actors. Alternative perspectives can be considered, recommendations can be challenged, and decisions can be justified independently of system outputs. In this configuration, algorithmic systems expand the informational environment without fundamentally restructuring the source of judgement. Human actors remain the primary authors of decisions, and responsibility largely corresponds to the authority exercised within the decision process.

5.2.2 Decision delegation

Decision delegation represents an intermediate configuration in which algorithmic outputs play a more central role in shaping decisions. Human actors continue to exercise oversight, yet portions of the decision process are routinely entrusted to AI systems because of perceived efficiency, consistency, or expertise. The practical scope of human discretion narrows as reliance on algorithmic recommendations becomes institutionalised. Although judgement remains formally shared, authority begins to migrate towards the system. Human actors increasingly function as validators of algorithmic outputs rather than as independent originators of judgement, creating a growing divergence between the formal allocation of responsibility and the practical distribution of decision-making authority.

5.2.3 Decision substitution

Decision substitution is likely to emerge when algorithmic outputs become the primary reference point through which decisions are understood, justified, and enacted. Human actors remain present within the process, but their role increasingly centres on validation, implementation, or procedural oversight. Judgement is no longer generated primarily through independent deliberation; it is organised around the outputs of the system. The defining feature of substitution is not the absence of human involvement but the displacement of human judgement as the principal source of decision authority. In this configuration, legitimacy increasingly attaches to algorithmic recommendations, while human actors retain accountability for outcomes despite exercising diminishing influence over the conditions through which those outcomes are produced.

This progression should not be interpreted as a linear or inevitable pathway. Organisations may move between configurations, and different aspects of the same decision process may exhibit varying degrees of support, delegation, and substitution. The typology should be understood as an analytical framework rather than a developmental sequence. Its value lies in illustrating how changing distributions of authority, discretion, and accountability reshape the relationship between human judgement and algorithmic systems. Figure 1 synthesises the central argument of the study. Rather than treating decision support, delegation, and substitution as isolated levels of automation, it illustrates how these configurations emerge from the broader transformation of the decision field by AI systems operating as cultural and epistemic infrastructures. As algorithmic systems increasingly shape visibility, validation, participation, timing, and legitimacy, authority progressively migrates from human actors towards algorithmic outputs. The result is a gradual displacement of human judgement that generates new forms of responsibility asymmetry and raises important societal and governance challenges.

The figure illustrates the research's central argument. AI systems operate as cultural and epistemic infrastructures that reorganise the decision field by shaping visibility, validation, timing, participation, and legitimacy. These

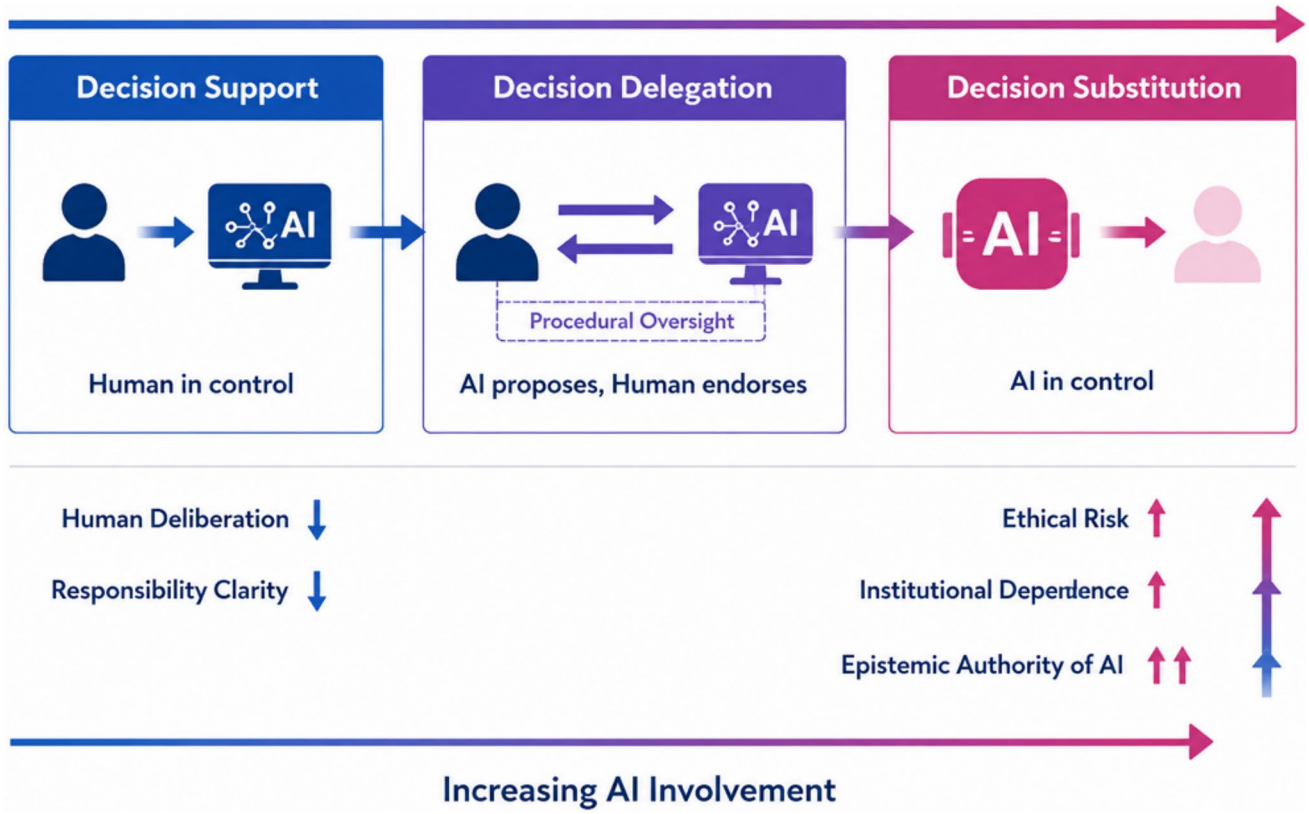

Fig. 1 From AI infrastructure to decision substitution: the progressive displacement of human judgement

transformations alter the relationship between human and algorithmic authority, producing different configurations of decision-making ranging from decision support and delegation to decision substitution. As judgement progressively shifts towards algorithmic systems, responsibility asymmetries, reduced contestability, epistemic narrowing, and governance challenges become increasingly salient. While the distinction between support, delegation, and substitution clarifies different configurations of human–AI decision-making, it does not by itself explain how or when movement towards substitution occurs. To address this issue, the analysis now identifies a set of threshold conditions through which algorithmic systems progressively acquire authority within the decision field, creating circumstances in which human judgement becomes increasingly constrained, displaced, or subordinated to system outputs.

5.3 Threshold conditions for decision substitution

A persistent challenge in discussions of AI-mediated decision-making is identifying the point at which decision support becomes decision substitution. The distinction is often treated as self-evident, yet in practice, the transition is rarely marked by a clear transfer of authority from humans to machines. Human actors frequently remain involved throughout the process, making it difficult to determine whether judgement has genuinely been displaced or merely augmented. The presence of human oversight alone is an insufficient indicator of meaningful agency. The argument advanced in this research is that decision substitution should not be defined by the absence of human participation, but by the transformation of the conditions under which judgement is exercised. Substitution becomes increasingly probable when algorithmic systems become the primary source of authority through which decisions are interpreted, justified, and enacted.

Building on the preceding analysis of the decision field, four threshold conditions can be identified. These conditions should not be understood as discrete stages or deterministic indicators. Rather, they represent interconnected shifts in the distribution of authority, legitimacy, and participation within the decision field. As these conditions become more pronounced, algorithmic systems increasingly acquire practical authority over decisions, while opportunities for independent human judgement become progressively constrained. They provide a framework for identifying when AI-mediated decision-making moves beyond support and delegation towards substitution.

5.3.1 The institutional cost of disagreement

The first threshold emerges when disagreement with algorithmic outputs becomes institutionally costly. In decision-support arrangements, human actors can challenge recommendations without incurring significant procedural or organisational penalties. Their judgement remains independently legitimate. As algorithmic systems become more deeply embedded, however, the burden of justification often shifts. Aligning with system outputs is treated as normal, efficient, and defensible, while deviation requires additional explanation, documentation, or managerial approval. Under such conditions, disagreement remains formally possible but becomes practically constrained (Pasquale 2015; Burrell 2016). Human actors may continue to possess the authority to override algorithmic recommendations, yet the institutional costs associated with doing so discourage meaningful dissent. The capacity to disagree survives in principle while diminishing in practice.

5.3.2 The default validity of algorithmic outputs

A second threshold occurs when algorithmic outputs become the default reference point against which alternative judgements are evaluated. In support configurations, recommendations contribute to deliberation alongside other forms of evidence and expertise. In substitution configurations, by contrast, the system's recommendation acquires presumptive validity (Yeung 2018). The significance of this shift lies in the redistribution of justification. Human actors no longer ask whether the system's recommendation is correct; instead, they must explain why it should not be followed. Authority migrates towards the algorithm because its outputs become the baseline from which reasoning proceeds. Judgement is increasingly exercised within the parameters established by the system rather than independently of them.

5.3.3 The reduction of meaningful corrective capacity

A third threshold concerns the ability of human actors to meaningfully alter decision outcomes. Human oversight is often presented as evidence that substitution has not occurred. Yet, oversight is only significant when it provides genuine opportunities for intervention. Where organisational workflows, technical complexity, time pressures, or institutional expectations limit the practical ability to challenge system outputs, corrective capacity becomes largely symbolic (Green and Chen 2019; Selbst et al. 2019). Under these circumstances, human actors may retain formal responsibility for decisions while exercising limited influence over their substantive direction. Oversight becomes procedural rather than deliberative. The human remains within the loop, but the loop no longer serves as a meaningful site of judgement.

5.3.4 The separation of responsibility from authority

The final threshold occurs when responsibility remains attached to human actors even as authority increasingly migrates towards algorithmic systems (Bovens 2007). This asymmetry represents one of the most significant consequences of AI-mediated decision-making. Individuals continue to be held accountable for outcomes, yet the conditions under which those outcomes are produced are increasingly shaped by infrastructures beyond their direct control. The result is a growing divergence between the capacity to influence decisions and the obligation to answer for them. Responsibility persists, but its relationship to authority becomes increasingly unstable. Human actors absorb the consequences of decisions while exercising diminishing control over the processes that generate them.

Taken individually, none of these conditions is sufficient to establish decision substitution. Together, however, they provide a framework for identifying when the practical authority of judgement has begun to migrate from human actors to algorithmic systems. Substitution should be understood not as a technological state but as a socio-technical condition characterised by constrained dissent, default algorithmic legitimacy, diminished corrective capacity, and asymmetrical responsibility (Pasquale 2015; Coeckelbergh 2020). This perspective clarifies why the mere presence of human oversight cannot be treated as evidence of meaningful human agency. What matters is not whether humans remain involved in decision-making processes, but whether they retain the capacity to shape, contest, and author the judgements that those processes produce.

5.4 Empirical observability and comparative application

Although the threshold conditions developed in Sect. 5.3 are conceptual rather than validated measurement constructs, they can guide empirical investigation of how decision substitution emerges across organisational and institutional settings. Their empirical value lies in directing attention towards observable arrangements through which authority, dissent, correction, and accountability are organised. Each condition may be examined through qualitative, quantitative, or comparative evidence without assuming that it can be reduced to a single universal metric. The institutional cost of disagreement may be observed through requirements for additional documentation, managerial approval, escalation procedures, professional sanctions, or increased exposure to retrospective scrutiny when actors depart from algorithmic recommendations. Relevant evidence could include override frequencies, reasons recorded for disagreement, time required to justify deviations, and interviews examining

whether professionals perceive dissent as institutionally supported or discouraged.

The default validity of algorithmic outputs may be assessed by examining where the burden of justification is located. Indicators may include the frequency with which recommendations are accepted without independent review, whether human actors must justify acceptance or only departure, and how organisational policies describe the evidential status of system outputs. Comparative analysis could determine whether algorithmic recommendations operate as one source of evidence among several or as the presumptively correct basis for action. Diminished corrective capacity may be examined through the practical opportunities available to inspect, revise, override, or appeal AI-mediated decisions. Relevant considerations include access to system information, time and expertise available for review, formal override permissions, escalation routes, and whether interventions meaningfully alter outcomes. The existence of a nominal override mechanism would not itself demonstrate corrective capacity if organisational or technical conditions prevent its effective use.

The separation of authority and accountability may be investigated by comparing who materially shapes a decision with who is formally required to justify its consequences. Organisational documents, workflow analyses, interviews, audit records, and accountability procedures may reveal whether responsibility remains concentrated upon front-line actors even when model design, institutional policy, or system-generated recommendations substantially determine outcomes. These indicators are not intended as a universal measurement scale or deterministic test. Their significance will vary across institutional, cultural, and sectoral contexts. They instead provide an empirical research agenda through which the proposed framework can be examined, refined, and compared across AI-mediated decision environments.

5.5 Illustrative vignette: decision substitution in recruitment screening

Consider an organisation that adopts an AI-assisted recruitment platform to support candidate shortlisting. The system analyses résumés, application forms, employment histories, and assessment results to generate rankings of applicants according to their predicted suitability for a role. Officially, hiring decisions remain the responsibility of human recruiters, and the system is presented as a tool designed to improve efficiency and consistency in candidate evaluation. Initially, the platform functions as a decision-support system. Recruiters review algorithmic rankings alongside their own assessments, considering contextual factors that may not be captured by the model. Candidates can be advanced despite lower rankings, and algorithmic recommendations remain one input among many. Human judgement continues to operate as the primary source of authority within the recruitment process. Over time, however, organisational reliance on the system increases. Recruiters are encouraged to prioritise candidates identified by the platform because doing so reduces processing time and aligns with performance targets. Managers begin to treat the system's recommendations as evidence of objective decision-making, while deviations from algorithmic rankings require additional explanation. Although recruiters retain formal authority to override recommendations, doing so becomes increasingly burdensome. The institutional cost of disagreement rises.

As the system becomes embedded within organisational routines, algorithmic rankings acquire default validity. Rather than asking whether the platform's recommendation should be accepted, recruiters increasingly focus on identifying reasons to depart from it. The burden of justification shifts from the system to the human decision-maker. What began as a recommendation gradually becomes the baseline against which all alternative judgements are evaluated. At the same time, meaningful corrective capacity diminishes. Recruiters may not have access to the full logic underlying algorithmic rankings, nor sufficient time to independently evaluate every application. Human oversight remains formally present, yet the practical ability to challenge recommendations becomes increasingly constrained. The recruitment process continues to involve human participation, but the substantive direction of decision-making is increasingly organised around system outputs.

The final threshold emerges when responsibility and authority become separated. If a candidate challenges a hiring decision or alleges unfair treatment, accountability rests with human recruiters and organisational leaders. Yet, the conditions shaping the decision, including candidate visibility, ranking criteria, and evaluation priorities, have been substantially structured by the AI system. Human actors remain responsible for outcomes while exercising diminishing control over the processes that produce those outcomes. In this scenario, decision substitution does not occur because recruiters disappear from the hiring process. Nor does it arise because the system makes decisions autonomously. Rather, substitution emerges through the cumulative effects of institutional reliance, default algorithmic legitimacy, constrained dissent, and diminished corrective capacity. Human actors remain present throughout the process, yet the practical authority of judgement has shifted towards algorithmic outputs. This example illustrates the central claim of the research: decision substitution is best understood not as the replacement of humans by machines, but as a transformation in the location and exercise of judgement. The critical question is not whether humans remain in the loop, but whether they continue to function as meaningful authors of the decisions for which they are held accountable.

6 Responsibility, indeterminacy, and the limits of algorithmic governance

As the preceding sections have shown, decision substitution emerges through the reorganisation of the decision field, the migration of epistemic authority, and the gradual displacement of human judgement. These transformations create profound challenges for responsibility, legitimacy, and governance. The central issue is not merely that AI systems influence decisions, but that authority increasingly migrates towards algorithmic infrastructures while responsibility remains attached to human actors. The analysis that follows examines how responsibility asymmetries emerge under conditions of decision substitution, why indeterminacy marks a critical boundary of automation, and what these dynamics reveal about the normative limits of algorithmic governance.

6.1 Responsibility asymmetries and moral ambiguity in AI-mediated decisions

As AI systems move from decision support towards substitution, a central societal consequence is the emergence of responsibility asymmetries. The preceding analysis demonstrated that decision substitution may occur when authority, legitimacy, and corrective capacity progressively migrate towards algorithmic systems while responsibility remains attached to human actors. Under these conditions, harm may occur even when no actor exercises complete control over the decision-making process. Responsibility does not disappear; rather, it becomes unevenly distributed across socio-technical arrangements in which those held accountable are not always those who possess meaningful influence over outcomes. This asymmetry lies at the heart of contemporary concerns regarding AI-mediated decision-making.

6.1.1 Distributed agency and the diffusion of responsibility

In AI-mediated decision-making, outcomes are often the result of interactions among developers, deployers, institutions, users, and technical systems. While such distribution is not unique to AI, algorithmic infrastructures intensify the problem by introducing opaque, probabilistic processes that mediate judgement without possessing moral agency themselves. When decisions are justified by reference to system outputs, risk scores, rankings, or generated recommendations, responsibility becomes fragmented and difficult to trace. Viewed through the lens of decision substitution, responsibility diffusion is not simply a consequence of technological complexity. It is a consequence of the separation of authority from accountability. As algorithmic systems increasingly shape what is visible, legitimate, and actionable within the decision field, human actors may remain formally responsible while exercising diminishing control over the substantive direction of decisions. This diffusion is experienced asymmetrically. Individuals at the point of decision are frequently held accountable for outcomes, even when their scope for deliberation has been significantly constrained by system design and institutional policy. Conversely, upstream actors, such as system designers or organisations, may invoke complexity, scale, or technical neutrality to deflect responsibility.

6.1.2 The illusion of control and procedural accountability

Many governance frameworks attempt to address responsibility asymmetries through procedural mechanisms, such as documentation requirements, audit trails, or human-in-the-loop mandates. While these measures are important, they often create an illusion of control rather than substantive accountability. The presence of a human actor in the loop does not guarantee meaningful agency if that actor lacks the authority, resources, or epistemic access required to contest system outputs. Empirical research shows that human oversight can become routinised and symbolic, particularly in high-volume or high-pressure environments. This dynamic corresponds directly to the erosion of meaningful corrective capacity identified in Sect. 5. Human oversight only functions as a safeguard when actors possess genuine opportunities to challenge, revise, or redirect algorithmic outputs. Where institutional pressures, workflow constraints, or epistemic dependence discourage such intervention, oversight becomes largely procedural. Human actors remain within the loop, yet their role increasingly resembles validation rather than judgement.

In high-stakes fields like healthcare, the presence of an AI recommendation can bias expert judgement, reducing oversight to a symbolic validation of algorithmic output (Gaube et al. 2021). This phenomenon mirrors longstanding findings in human factors research, where operators under cognitive load become complacent and unduly reliant on automation (Parasuraman and Manzey 2010). In bureaucratic settings, this can render human review a ceremonial "rubber-stamp" rather than a substantive check, as seen in the algorithmic governance of social media content and judicial decisions (Gorwa et al. 2020; Stevenson and Doleac 2022). In such cases, accountability is formally assigned but practically hollowed out, reinforcing moral ambiguity rather than resolving it.

6.1.3 Moral ambiguity without moral agents

A recurring response to responsibility asymmetries is the temptation to attribute moral or quasi-agentive status to AI

systems themselves. However, treating AI as a moral agent risks obscuring the fundamentally human and institutional origins of algorithmic decision-making (Moser et al. 2022). AI systems do not possess intentions, moral understanding, or the capacity for accountability; their apparent agency is derivative of social practices and design choices. This misattribution contributes to moral ambiguity, where outcomes are perceived as inevitable consequences of "what the system decided." Such narratives naturalise algorithmic authority and diminish the perceived responsibility of those who design, deploy, and rely upon these systems.

6.1.4 Structural conditions producing responsibility asymmetries

Responsibility asymmetries should be understood not as accidental failures, but as structural features of AI-mediated governance. These asymmetries are reinforced by the structural, functional, and normative reorganisation of the decision field. As AI systems increasingly shape visibility, validation criteria, decision timing, participation, and institutional legitimacy, authority progressively migrates towards algorithmic infrastructures even while accountability remains attached to individuals and organisations. The result is a growing divergence between those who influence decisions and those who ultimately answer for their consequences. Responsibility asymmetries emerge not simply because decision-making is shared between humans and machines, but because the practical authority to shape judgements becomes increasingly separated from the obligation to justify and account for their outcomes.

They emerge from (1) institutional incentives favouring efficiency and risk aversion, (2) epistemic asymmetries between system designers and users, (3) cultural narratives of technological objectivity, and (4) regulatory frameworks that lag behind socio-technical change. These conditions encourage decision substitution while preserving the appearance of human oversight. As a result, responsibility becomes increasingly difficult to locate and contest: it is distributed across complex socio-technical arrangements, yet frequently concentrated upon those with the least capacity to influence the systems shaping outcomes (Suri 2025). The challenge is not the disappearance of responsibility, but its redistribution in ways that separate accountability from practical authority within AI-mediated decision processes.

6.1.5 Implications for ethical and legal frameworks

The persistence of responsibility asymmetries challenges prevailing ethical and legal approaches to AI governance. Frameworks focussed on transparency, explainability, or individual accountability are insufficient if they do not address how agency is practically exercised and constrained within institutional contexts (Floridi et al. 2018). Without confronting decision substitution directly, such approaches risk legitimising systems that erode human judgement while shielding powerful actors from scrutiny. Responsibility asymmetries persist not only because decision-making is distributed across multiple actors, but also because legitimacy increasingly attaches to algorithmic outputs themselves. As decisions become defensible by virtue of their alignment with system recommendations, the practical authority of human judgement may erode even where formal accountability structures remain unchanged. Ethical and legal frameworks must address not only who is accountable, but also how authority is produced, distributed, and legitimised within AI-mediated environments. The following section builds on this analysis by examining the role of indeterminacy and exclusion in AI-mediated decision-making, and why responsibility asymmetries are most acute precisely where algorithmic systems are least equipped to represent what matters.

6.2 Indeterminacy, exclusion, and the limits of algorithmic representation

A defining feature of AI-mediated decision-making is that it operates through formalisation: transforming aspects of the world into variables, categories, and optimisation targets. While this capacity enables scale and efficiency, it also introduces a fundamental limitation. AI systems can only act upon what is rendered legible to them, what can be measured, classified, or learned from historical data. This creates a structural tension between algorithmic decision-making and forms of social judgement that depend on indeterminacy, tacit knowledge, and contextual interpretation. From a socio-technical perspective, indeterminacy refers to those aspects of human life that cannot be fully specified in advance: moral ambiguity, cultural meaning, relational nuance, and situations where values conflict rather than align (Bowker and Star 1999). Such conditions are not anomalies to be eliminated; they are constitutive of social life. Yet, AI systems are typically designed to reduce indeterminacy, producing determinate outputs even where uncertainty, contestation, or plurality would be more appropriate (Kitchin 2019).

This mismatch becomes ethically consequential when decision substitution occurs in precisely those contexts where indeterminacy is most salient. Algorithmic systems may offer rankings, scores, or generated narratives that appear authoritative while obscuring the value judgements embedded in their construction (Alon-Barkat and Busuioc 2023). In doing so, they risk foreclosing alternative interpretations and marginalising forms of reasoning that do not conform to dominant data regimes. The significance of indeterminacy becomes particularly apparent once decision

substitution has occurred. Where algorithmic systems increasingly define what counts as relevant evidence, acceptable reasoning, and legitimate action, the limits of representation become governance concerns rather than merely technical concerns. The challenge is no longer simply whether AI systems can model reality accurately, but whether certain dimensions of human judgement ought to remain resistant to formalisation altogether.

6.2.1 Exclusion through representation

Because AI systems rely on historical data and predefined categories, they systematically privilege certain experiences and perspectives while excluding others. Forms of knowledge that are relational, embodied, or culturally situated are often underrepresented or distorted, leading to what has been described as representational harm. These exclusions are not merely technical shortcomings; they reflect broader social inequalities and power relations encoded into data infrastructures. Empirical research demonstrates that marginalised populations are disproportionately affected by such exclusions, particularly in domains such as welfare allocation, risk assessment, and automated screening. Landmark investigations have shown how algorithmic systems in criminal justice perpetuate racial bias, falsely flagging Black defendants as high risk at disproportionate rates (Angwin et al. 2016). In public services, automated welfare and fraud detection systems function as "weapons of math destruction," profiling and punishing the poor while draining the resources needed to contest erroneous decisions (O'Neil 2016). These harms extend to economic and health opportunities, where biased hiring algorithms screen out marginalised applicants and clinical algorithms systematically deprioritise Black patients for crucial care by embedding historic inequities into their logic (Raghavan et al. 2020; Obermeyer et al. 2019). When algorithmic outputs substitute for human judgement, these exclusions become institutionalised, rendering them more difficult to contest or even recognise.

6.2.2 Epistemic narrowing and the loss of contestability

A further consequence of decision substitution is epistemic narrowing, the contraction of what counts as relevant knowledge or legitimate reasoning within decision processes. As AI systems standardise forms of evidence and justification, alternative perspectives may be dismissed as anecdotal, subjective, or inefficient (Seaver 2019). From the perspective of the decision field developed earlier in this research, epistemic narrowing also weakens the conditions necessary for meaningful dissent. When algorithmic outputs become default-valid and alternative forms of reasoning are increasingly marginalised, contestability shifts from a practical possibility to a formal entitlement. Individuals may retain the right to challenge decisions while lacking the institutional conditions necessary for such challenges to be effective.

Over time, this can erode spaces for disagreement and deliberation, replacing them with outputs that appear neutral but embody specific normative commitments. From the perspective of philosophy of technology, this dynamic reflects a shift in moral mediation. Technologies shape not only actions, but also the horizons of interpretation within which ethical reasoning takes place (Verbeek 2024). When AI systems dominate decision-making, they implicitly define which questions can be asked and which answers are admissible, thereby constraining moral imagination.

6.2.3 Indeterminacy as a site of human responsibility

Recognising the limits of algorithmic representation does not entail rejecting AI systems outright. Rather, it underscores the need to preserve human judgement precisely where indeterminacy persists. Ethical responsibility is most acute not in routine or well-defined cases, but in situations requiring sensitivity to context, empathy, and value pluralism. This view aligns with care ethics, which locates morality in the responsive, particular, and relational dimensions of human life that resist codification (Tronto 1993). Techno-moral virtue theorists similarly argue that ethical expertise rests in phronesis, practical wisdom for navigating ambiguity, a capacity that remains beyond algorithmic replication (Vallor 2016).

Within socio-technical systems, this irreducible residue of human judgement is also where responsibility becomes concentrated, often producing what Elish (2019) terms "moral crumple zones," in which human actors absorb blame when value conflicts arise within automated systems. While AI may effectively manage tractable routines, the burden of ethical decision-making intensifies precisely where context, empathy, and pluralistic values are paramount. Decision substitution in such domains, therefore, risks displacing the very capacities that make moral and social judgement possible. This insight reinforces the research's central claim: the ethical challenge posed by AI lies less in machine intelligence itself than in how societies govern the boundary between automation and human responsibility. When AI systems are permitted to substitute for judgement in indeterminate domains, exclusion and responsibility asymmetries become systemic rather than exceptional.

6.2.4 Implications for AI-mediated futures

By foregrounding indeterminacy and exclusion, this section underscores why the societal risks of decision substitution cannot be addressed through technical fixes alone. Transparency, accuracy, and fairness metrics are insufficient if they do not confront the question of what cannot, and should not,

be automated. A society-shaping-the-algorithm approach requires recognising the ethical value of indeterminacy and designing governance structures that protect spaces for human deliberation, contestation, and care. The next section turns directly to these governance implications, revisiting the notion of human-in-the-loop and proposing pathways towards participatory, accountable, and value-sensitive AI-mediated decision-making.

6.3 Human-in-the-loop revisited: from procedural oversight to societal governance

The preceding sections have shown that decision substitution emerges through the reorganisation of the decision field, the migration of epistemic authority, and the separation of responsibility from authority. These dynamics reveal important limitations in prevailing interpretations of human-in-the-loop governance. Although human oversight is frequently presented as a safeguard against excessive automation, the mere presence of a human actor does not guarantee meaningful agency. Human-in-the-loop arrangements can coexist with decision substitution when oversight becomes procedural, corrective capacity is constrained, and algorithmic outputs acquire default legitimacy. In this context, human-in-the-loop should be understood not merely as a technical design principle but as a broader societal governance challenge. Figure 2 illustrates this shift by contrasting procedural human-in-the-loop arrangements with a societal governance perspective, emphasising that meaningful oversight requires collective, ex ante, and value-driven forms of human involvement rather than symbolic compliance.

Figure 2 contrasts procedural interpretations of human-in-the-loop (HITL), which emphasise individual oversight, post hoc checks, and compliance, with a societal governance approach that foregrounds collective responsibility, ex ante deliberation, and legitimacy. The reframing highlights that meaningful human involvement in AI-mediated decision-making depends not on the mere presence of a human actor, but on institutional conditions that enable participation, contestation, and value-sensitive judgement.

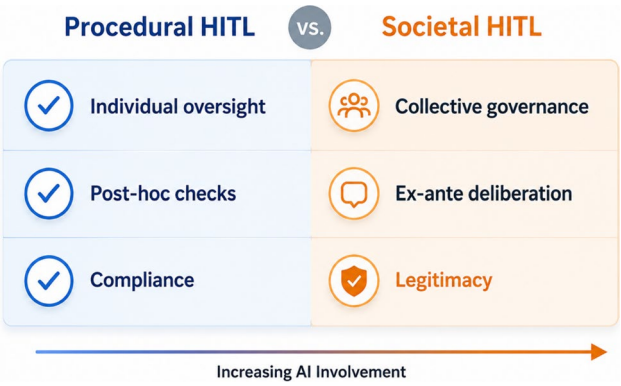

Fig. 2 Reframing human-in-the-loop governance from procedural oversight to societal governance

6.3.1 The limits of procedural human-in-the-loop

In many AI governance frameworks, human-in-the-loop is defined minimally as the presence of a human actor who can review, approve, or override algorithmic outputs (Santoni de Sio and Van den Hoven 2018; Shneiderman 2020). While appealing in principle, empirical research shows that such arrangements frequently devolve into symbolic oversight. Human actors may lack the time, authority, or epistemic access required to meaningfully contest system outputs, particularly in high-volume or high-stakes settings (Beede et al. 2020; Burrell 2016). Under these conditions, the "loop" risks becoming a rubber-stamping mechanism, preserving the appearance of accountability while leaving decision substitution intact. Responsibility is formally assigned to humans, yet substantively shaped by system design and institutional pressure. This disconnect contributes to moral ambiguity rather than ethical clarity.

6.3.2 Human judgement as a collective and institutional practice

Reframing human-in-the-loop requires recognising that judgement is not an isolated cognitive act but a collective, institutional, and relational practice. Decisions are shaped by organisational norms, performance metrics, professional cultures, and regulatory expectations (Waggoner et al. 1999; Eubanks 2018). Ensuring meaningful human involvement, therefore, demands attention to these broader conditions, not merely interface design. From this perspective, human-in-the-loop governance must address: (1) who is empowered to challenge algorithmic outputs, (2) under what conditions dissent is legitimate, and (3) how institutional incentives shape reliance on AI systems. Without such structural support, appeals to human oversight risk placing disproportionate moral burden on individuals while leaving systemic drivers of decision substitution unexamined.

6.3.3 Institutional conditions for meaningful human intervention

Meaningful human involvement requires more than assigning a person formal authority to approve or override an algorithmic output. It depends upon institutional conditions that make intervention practically possible, epistemically informed, and professionally legitimate. Without these conditions, human-in-the-loop arrangements may retain the appearance of oversight while leaving the distribution of decision authority substantially unchanged. First, human

actors require sufficient time, resources, and relevant expertise to evaluate system outputs rather than merely process them. Oversight conducted under severe workload, performance, or temporal pressures is unlikely to support independent deliberation. Second, reviewers must have access to information proportionate to their responsibilities, including the basis, uncertainty, limitations, and intended scope of algorithmic recommendations. Meaningful intervention does not require complete technical transparency, but it does require sufficient epistemic access to identify when reliance may be inappropriate.

Third, institutions must provide genuine corrective authority. Human actors should be able to revise, override, suspend, or escalate AI-mediated decisions, and these interventions must be capable of changing substantive outcomes. Fourth, justified dissent must be institutionally protected. Where disagreement produces excessive documentation burdens, professional sanctions, or heightened exposure to blame, the formal capacity to intervene will remain practically weak. Override and appeal processes should be treated as legitimate components of responsible practice rather than as deviations from efficient system use. Fifth, organisations require mechanisms for collective learning. Overrides, errors, contested decisions, and stakeholder appeals should inform system review, institutional policy, and future deployment rather than remaining isolated acts of individual correction. Finally, authority and accountability should be aligned as closely as possible. Actors should not bear full responsibility for decisions when they lack meaningful influence over system selection, design, deployment conditions, or operational use. These conditions distinguish human presence from institutionally supported human agency. Human-in-the-loop governance is meaningful only where individuals and affected communities possess credible opportunities to understand, contest, correct, and reshape AI-mediated decisions without disproportionate institutional cost.

6.3.4 Participatory and value-sensitive governance

Consistent with the human-centred tradition, a more robust approach to governance emphasises participation, reflexivity, and value-sensitivity. Participatory design and governance practices invite affected stakeholders, not only developers or managers, but also users and impacted communities, into decisions about how AI systems are designed, deployed, and evaluated. Such approaches shift the focus from post hoc oversight to ex ante deliberation about acceptable roles for AI in specific social contexts. Rather than asking only whether a system is accurate or efficient, participatory governance foregrounds questions of legitimacy, inclusion, and care, questions that cannot be resolved through optimisation alone.

6.3.5 Making the role of AI explicit

A key implication of the decision support–delegation–substitution framework is the need for explicit declaration of AI's role within decision processes. Institutions should be required to specify whether AI systems are intended to support human judgement, delegate provisional decisions, or substitute for deliberation altogether. Such transparency enables more meaningful public scrutiny and aligns accountability with actual practice rather than aspirational framing. Explicit role definition also helps protect spaces of indeterminacy by identifying domains in which substitution is ethically inappropriate. In doing so, governance shifts from reactive control to proactive boundary-setting.

6.3.6 Towards society shaping the algorithm

Revisiting human-in-the-loop in this way supports a broader reorientation from algorithmic governance shaping society to society shaping the algorithm. This requires moving beyond narrow compliance metrics towards governance frameworks that recognise AI systems as socio-technical actors embedded in cultural and institutional life (Beer 2019). It also entails acknowledging that some forms of judgement, particularly those involving moral ambiguity, relational care, or value conflict, should remain irreducibly human. The final section synthesises these arguments, reflecting on their implications for AI-mediated futures and outlining directions for research and practice that preserve human agency, responsibility, and social meaning.

7 Discussion: from algorithmic governance to society shaping the algorithm

This study set out to examine a largely unacknowledged transformation in AI-mediated life: the shift from decision support to decision substitution and its consequences for human agency, responsibility, and social meaning. The preceding analysis demonstrates that this shift is not merely a technical evolution, but a societal reconfiguration of judgement, in which AI systems increasingly function as epistemic and moral infrastructures rather than neutral tools. More specifically, the analysis has shown that AI systems increasingly reorganise the decision field itself. By shaping visibility, validation, participation, timing, and legitimacy, AI systems influence not only what decisions are made but also the conditions under which decisions become possible. Decision substitution emerges within this broader transformation, making it necessary to move beyond narrow accounts of automation and towards a richer understanding of how judgement is socially, institutionally, and technologically organised.

7.1 Reframing the ethical problem of AI

A central contribution of this study is to reframe the ethical challenge of AI away from isolated concerns such as bias, transparency, or explainability, towards the distribution and displacement of human judgement. While existing AI ethics frameworks have generated important normative principles, they often presuppose that humans remain the primary decision-makers (Floridi et al. 2018). The analysis developed in this research challenges this assumption by demonstrating that the central ethical issue is not automation alone, but the reorganisation of the decision field through which judgement is exercised. The progression from support and delegation to substitution reflects a broader transformation in which algorithmic systems increasingly shape the structural, functional, and normative conditions of decision-making. Human agency may be diminished not because humans disappear from the process, but because the practical authority of judgement progressively migrates towards algorithmic outputs.

This reframing aligns with longstanding critiques of solutionist and optimisation-driven approaches to technology. The ethical question is not only whether AI systems perform well, but also whether they reorganise decision-making in ways that undermine deliberation, contestability, and care. From this perspective, ethical concerns such as bias, transparency, and explainability remain important but become insufficient as standalone frameworks. The deeper question concerns how AI systems reshape the institutional environments within which authority is exercised, legitimacy is produced, and responsibility is allocated.

7.2 Decision substitution as a societal phenomenon

By conceptualising AI systems as cultural and epistemic infrastructures, this study situates decision substitution within broader processes of institutionalisation and normalisation. The concept of decision field reorganisation helps explain how this process occurs. Rather than directly replacing human actors, AI systems progressively alter the positions, functions, and sources of legitimacy through which decisions are produced. Structural transformations redistribute authority, functional transformations shape visibility and validation, and normative transformations redefine what counts as a reasonable or defensible judgement. Decision substitution emerges not through explicit intent, but through everyday practices shaped by organisational incentives, epistemic authority, and time pressure. This helps explain why responsibility asymmetries persist even in systems nominally designed to keep humans "in the loop." Human actors frequently remain accountable for outcomes while exercising diminishing influence over the conditions through which those outcomes are produced. The result is a growing divergence between authority and accountability that existing governance frameworks struggle to address.

7.3 Power, exclusion, and the politics of substitution

The discussion also foregrounds the political dimensions of decision substitution. Empirical literature shows that marginalised groups are more likely to encounter AI systems as instruments of surveillance, classification, and control, while having fewer opportunities to contest or reinterpret their outputs (Eubanks 2018). The analysis presented here suggests that these inequalities are reinforced through the reorganisation of the decision field itself. When algorithmic systems increasingly determine what becomes visible, whose experiences are recognised, and which forms of knowledge are treated as legitimate, power becomes embedded within infrastructural arrangements that are often difficult to contest or even perceive.

Decision substitution, therefore, reinforces existing power asymmetries, embedding them within ostensibly neutral infrastructures. From this perspective, substitution is not only an ethical issue, but also a governance and justice issue. Addressing it requires confronting whose values are prioritised, whose knowledge is rendered legible, and whose agency is constrained by algorithmic systems. Decision substitution raises questions not only about individual decisions but also about the distribution of epistemic authority across society. The issue is not merely whether AI systems make fair decisions, but whose perspectives are incorporated into the infrastructures that shape judgement itself.

7.4 Rethinking human-in-the-loop

The study's critique of procedural human-in-the-loop approaches further underscores the need for a societal rather than technical understanding of governance. As shown in earlier sections, the mere presence of a human actor does not guarantee meaningful oversight if institutional conditions favour deference to algorithmic outputs (Parasuraman et al. 2000). The threshold conditions developed earlier in the work demonstrate that meaningful oversight depends not simply on the presence of a human actor, but on the preservation of corrective capacity, the legitimacy of dissent, and the continued authority of human judgement. Where these conditions are absent, human participation may remain visible while becoming substantively weakened.

Human-in-the-loop must be reconceptualised as a collective, participatory, and value-sensitive practice rather than an individual compliance mechanism. This reconceptualisation supports a shift from algorithmic governance shaping society towards society shaping the algorithm. Such a shift requires governance approaches capable of influencing not

only algorithmic outputs, but also value commitments, and legitimacy structures through which AI systems acquire authority. In addition, it requires not only new technical designs but also new institutional arrangements that protect spaces for deliberation, dissent, and ethical reflection.

7.5 Implications for research and practice

The framework developed in this work offers a foundation for future interdisciplinary research into AI-mediated decision-making. It extends beyond existing debates on algorithmic fairness or transparency by providing a conceptual vocabulary for analysing how AI systems reorganise the conditions of judgement. In particular, the concepts of decision field reorganisation, decision substitution, and responsibility asymmetry offer new avenues for examining how authority, legitimacy, and accountability are redistributed across socio-technical systems. Empirically, it invites qualitative and comparative studies of how decision substitution is experienced across sectors, cultures, and institutional contexts. Normatively, it calls for governance approaches that explicitly delimit where and when AI should not substitute for human judgement, and how institutional arrangements can preserve meaningful opportunities for deliberation, contestation, and corrective intervention.

More broadly, the analysis suggests that preserving human agency in AI-mediated futures is not a matter of resisting technology, but of actively shaping its role in social life. This requires recognising that some dimensions of judgement, particularly those involving moral ambiguity, relational care, and value conflict, are not failures of computation, but expressions of what it means to be human. Hence, the significance of decision substitution lies not in the prospect of machines replacing humans, but in the gradual transformation of the conditions under which human judgement is exercised. AI systems increasingly function as infrastructures that shape visibility, legitimacy, participation, and authority across diverse domains of social life. Understanding these transformations requires moving beyond questions of technical performance towards a broader examination of how societies organise judgement, distribute responsibility, and determine which forms of reasoning remain meaningful in increasingly AI-mediated futures.

8 Conclusion: reclaiming judgement in AI-mediated futures

This study examines the ethical, societal, and philosophical implications of a subtle yet consequential transformation in contemporary AI use: the shift from decision support to decision substitution. By conceptualising AI systems as cultural and epistemic infrastructures and examining how they reorganise the decision field, the analysis has shown that the central risk of AI-mediated decision-making lies not solely in error or bias, but in the progressive transformation of the conditions under which judgement is exercised. The study has argued that decision substitution emerges when algorithmic systems increasingly shape visibility, validation, participation, and legitimacy, creating environments in which human actors remain formally involved while the practical authority of judgement progressively migrates towards algorithmic outputs. Across everyday institutional and personal contexts, AI systems increasingly shape how decisions are framed, justified, and enacted. These systems do not simply automate tasks; they reorganise positions, functions, and sources of legitimacy within decision-making processes. In doing so, they reshape how authority is exercised, how competence is defined, and how decisions become justified and actionable in institutional and everyday settings. As a result, responsibility asymmetries emerge, particularly where human actors remain accountable for outcomes while exercising diminishing influence over the socio-technical conditions through which those outcomes are produced (Danaher 2022).

A key contribution of this work is to demonstrate that decision substitution is most ethically problematic in domains characterised by indeterminacy, where moral ambiguity, contextual understanding, and value pluralism are unavoidable. The analysis demonstrates that these limitations become especially significant once decision substitution has occurred. Where algorithmic outputs acquire default legitimacy and opportunities for contestation become constrained, the capacity of human actors to exercise interpretive judgement may be progressively weakened even when formal oversight mechanisms remain in place. In such settings, algorithmic optimisation risks crowding out precisely those forms of judgement that sustain social meaning, democratic deliberation, and care. Recognising the limits of algorithmic representation is not a rejection of AI, but an affirmation of the ethical significance of human interpretation, contestation, and responsibility. Some dimensions of judgement cannot be reduced to optimisation problems because they involve negotiating uncertainty, balancing competing values, and responding to contexts that resist formal representation.

The findings further argue that prevailing governance approaches, including procedural interpretations of human-in-the-loop, are insufficient to address these challenges. Meaningful oversight cannot be reduced to technical checkpoints or individual accountability. Instead, it requires societal governance arrangements that make the role of AI in decision-making explicit, protect spaces for human deliberation, and enable participation by those affected by algorithmic systems. This analysis supports a shift from algorithmic governance shaping society to society shaping the algorithm. Achieving this shift requires recognising that AI systems

are not merely tools deployed within social environments, but infrastructures that increasingly shape how knowledge, authority, and legitimacy are organised. Governance must address not only system outputs, but also the institutional and cultural conditions through which algorithmic authority is produced and sustained.

Such a shift demands attention not only to system design, but also to the cultural, institutional, and normative contexts in which AI is embedded. It also calls for future research that empirically examines lived experiences of AI-mediated decision-making across diverse settings and cultures, and that explores governance models capable of preserving human agency in the face of increasing automation. Future research should also examine how the reorganisation of the decision field varies across domains, cultures, and institutional contexts, and explore mechanisms capable of preserving meaningful opportunities for dissent, corrective intervention, and collective participation in AI-mediated decision-making. The challenge posed by AI is not whether machines can decide, but how societies choose to organise the conditions under which judgement remains meaningful. As AI systems become increasingly embedded within the infrastructures of everyday life, preserving human agency will depend less on maintaining symbolic human oversight and more on sustaining the institutional, cultural, and normative conditions that enable deliberation, contestation, and responsibility. Reclaiming those conditions is essential if AI is to contribute to human flourishing rather than quietly erode the foundations upon which ethical, democratic, and social life depend.

Author contributions AOI: Conceptualisation; Methodology; Theoretical framework development; Writing—original draft; Writing—review & editing; Supervision. OA: Literature synthesis; Interdisciplinary analysis (STS and governance perspectives); Writing—review & editing. DDA: legal, Empirical literature analysis; Human–AI interaction and institutional practice review; Writing—review & editing. RSAK: Governance and accountability analysis; Critical revisions; Writing—review & editing. RFF: Responsibility and ethical framework analysis; Writing—review & editing.

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

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alkhatib A, Bernstein MS (2019) Street-level algorithms: a theory at the gaps between policy and decisions. Proc CHI Conf Hum Factors Comput Syst 530:1–13. https://doi.org/10.1145/3290605.3300760

Alon-Barkat S, Busuioc M (2023) Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. J Public Adm Res Theory 33(1):153–169. https://doi.org/10.1093/jopart/muac035

Amoore L (2020) Cloud ethics: algorithms and the attributes of ourselves and others. Duke University Press

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Retrieved January 10, 2026.

Arendt H (2019) The human condition. University of Chicago press

Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, Vardoulakis LM (2020) A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Proceedings of the 2020 CHI conference on human factors in computing systems. Association for Computing Machinery, pp 1–12. https://doi.org/10.1145/3313831.3376718

Beer D (2019) The social power of algorithms. The social power of algorithms. Routledge, pp 1–13. https://doi.org/10.4324/9781351200677

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, pp 610–623. https://doi.org/10.1145/3442188.3445922

Benjamin R (2023) Race after technology. Social theory re-wired. Routledge, pp 405–415

Berger P, Luckmann T (2016) The social construction of reality. Social theory re-wired. Routledge, pp 110–122. https://doi.org/10.4324/9781315775357

Bijker WE, Law J (eds) (1994) Shaping technology/building society: studies in sociotechnical change. MIT Press

Birhane A, Kasirzadeh A, Leslie D, Wachter S (2023) Science in the age of large language models. Nat Rev Phys 5(5):277–280. https://doi.org/10.1038/s42254-023-00571-z

Boddu Y (2025) AI agents for real-time operational decision making: embedding intelligence in data pipelines. J Comput Sci Technol Stud 7(5):867–872

Bovens M (2007) Analysing and assessing accountability: a conceptual framework. Eur Law J 13(4):447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bowker GC, Star SL (1999) Sorting things out: classification and its consequences. MIT Press. https://doi.org/10.7551/mitpress/6352.001.0001

Brey PA (2005) Artifacts as social agents. In: Harbers H (ed) Inside the politics of technology: agency and normativity in the production of technology and society. Amsterdam University Press

Burrell J (2016) How the machine "thinks": understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512

Cabitza F, Rasoini R, Gensini GF (2017) Unintended consequences of machine learning in medicine. JAMA 318(6):517–518. https://doi.org/10.1001/jama.2017.7797

Chiodo S (2023) Technology and the overturning of human autonomy. Springer. https://doi.org/10.1007/978-3-031-26159-6
Christin A (2017) Algorithms in practice: comparing web journalism and criminal justice. Big Data Soc. https://doi.org/10.1177/2053951717718855
Coeckelbergh M (2020) Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 26:2051–2068. https://doi.org/10.1007/s11948-019-00146-8
Couldry N, Hepp A (2018) The mediated construction of reality. Wiley
Danaher J (2022) Tragic choices and the virtue of techno-responsibility gaps. Philos Technol 35(2):26. https://doi.org/10.1007/s13347-022-00519-1
de la Torre PG, Pérez-Verdugo M, Barandiaran XE (2026) Attention is all they need: cognitive science and the (techno) political economy of attention in humans and machines. AI Soc 41(1):5–21. https://doi.org/10.1007/s00146-025-02400-z
Deroncele-Acosta A, Gross-Tur R, Bellido-Valdiviezo O, López-Mustelier R (2024) Qualitative research from grounded theory to build a scientific framework on the researcher's epistemic competence. Int J Qual Methods. https://doi.org/10.1177/16094069241284218
Elish MC (2019) Moral crumple zones: cautionary tales in human–robot interaction. Engag Sci Technol Soc 5:40–60. https://doi.org/10.17351/ests2019.260
Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press
Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5
Fricker M (2017) Evolving concepts of epistemic injustice. The Routledge handbook of epistemic injustice. Routledge, pp 53–60. https://doi.org/10.4324/9781315212043
Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Ghassemi M (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 4(1):Article 31. https://doi.org/10.1038/s41746-021-00407-5
Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127. https://doi.org/10.1136/amiajnl-2011-000089
Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc. https://doi.org/10.1177/2053951720912992
Green B, Chen Y (2019) Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In: Proceedings of the conference on fairness, accountability, and transparency, pp 90–99. https://doi.org/10.1145/3287560.3287563
Hosseini SH, Sakhaei S (2025) Educating intelligence, producing power: Iranian sociologists on AI, knowledge production, and global hierarchies. J World Sociopolit Stud 9(4):887–921
Ibitoye AO, Ravindran R, Afe OF, Abiodun AO (2025a) Predicting and mitigating digital addiction with machine learning models for improved mental health outcomes. J Soc Comput 6(4):359–377. https://doi.org/10.23919/JSC.2025.0020
Ibitoye AO, Nkwo MS, Orji R (2025b) Rethinking responsible AI from ethical pillars to sociotechnical practice. AI Ethics 5(6):6207–6223. https://doi.org/10.1007/s43681-025-00809-2
Ihde D (1990) Technology and the lifeworld: from garden to earth. Indiana University Press. https://doi.org/10.2307/j.ctt16gz5jh
Jakesch M, Hancock JT, Naaman M (2023) Human heuristics for AI-generated language are flawed. Proc Natl Acad Sci U S A 120(11):e2208839120. https://doi.org/10.1073/pnas.2208839120
Jasanoff S (2004) States of knowledge: the co-production of science and social order. Routledge. https://doi.org/10.4324/9780203413845
Kazim T, Tomlinson J (2023) Automation bias and the principles of judicial review. Jud Rev 28(1):9–16. https://doi.org/10.1080/10854681.2023.2189405
Kellogg KC, Valentine MA, Christin A (2020) Algorithms at work: the new contested terrain of control. Acad Manag Ann 14(1):366–410. https://doi.org/10.5465/annals.2018.0174
Kelly M (2025) Situated epistemic infrastructures: a diagnostic framework for post-coherence knowledge. https://arxiv.org/abs/2508.04995
Khurana S (2025) Decolonizing artificial intelligence: indigenous knowledge systems, epistemic pluralism, and the ethics of technology. J Comput Allied Intell 3(3):1–10
Kitchin R (2019) Thinking critically about and researching algorithms. The social power of algorithms. Routledge, pp 14–29. https://doi.org/10.4324/9781351200677
Langer C (2024) Decision-making power and responsibility in an automated administration. Discov Artif Intell 4(1):59. https://doi.org/10.1007/s44163-024-00152-1
Latour B (2005) Reassembling the social: an introduction to actor-network-theory. Oxford University Press. https://doi.org/10.1093/oso/9780199256044.001.0001
Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. MIS Q 45(3):1501–1525. https://doi.org/10.25300/MISQ/2021/16559
Lehto MR, Nanda G, Nanda G (2021) Decision-making models, decision support, and problem solving. Handbook of human factors and ergonomics. Wiley, pp 159–202. https://doi.org/10.1002/9781119636113.ch6
Luusua A (2023) Katherine Crawford: Atlas of AI: power, politics, and the planetary costs of artificial intelligence. AI Soc 38:1257–1259. https://doi.org/10.1007/s00146-022-01488-x
Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679
Moser C, Den Hond F, Lindebaum D (2022) Morality in the age of artificially intelligent algorithms. Acad Manag Learn Educ 21(1):139–155. https://doi.org/10.5465/amle.2020.0287
Noble SU (2018) Algorithms of oppression: how search engines reinforce racism. NYU Press
Norman DA (2013) The design of everyday things, revised and expanded ed. Basic Books. https://doi.org/10.1201/9780429464245
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342
Orlikowski WJ (2007) Sociomaterial practices: exploring technology at work. Organ Stud 28(9):1435–1448. https://doi.org/10.1177/0170840607081138
Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52(3):381–410. https://doi.org/10.1177/0018720810376055
Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. Hum Factors 39(2):230–253
Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A 30(3):286–297. https://doi.org/10.1109/3468.844354
Pasquale F (2015) The black box society: the secret algorithms that control money and information. Harvard University Press

Perrotta C, Selwyn N (2020) Deep learning goes to school: toward a relational understanding of AI in education. Learn Media Technol 45(3):251–269. https://doi.org/10.1080/17439884.2020.1686017

Popow M (2026) Decolonizing knowledge in the postdigital era: pedagogical strategies for navigating AI-driven epistemic transformations. Educ Philos Theory 58(1–2):44–56. https://doi.org/10.1080/00131857.2025.2552764

Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, pp 469–481. https://doi.org/10.1145/3351095.3372828

Rahwan I, Cebrian M, Obradovich N et al (2019) Machine behaviour. Nature 568:477–486. https://doi.org/10.1038/s41586-019-1138-y

Ricoeur P (1992) Oneself as another. University of Chicago press

Romeo G, Conti D (2025) Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. AI Soc. https://doi.org/10.1007/s00146-025-02422-7

Rosa H (2013) Social acceleration: a new theory of modernity. Columbia University Press

Santoni de Sio F, Van den Hoven J (2018) Meaningful human control over autonomous systems: a philosophical account. Front Robot AI 5:Article 15. https://doi.org/10.3389/frobt.2018.00015

Seaver, N. (2019). Knowing algorithms. In J. Vertesi & D. Ribes (Eds.), digitalSTS: A field guide for science & technology studies (pp. 412–422). Princeton University Press. https://doi.org/10.2307/j.ctvc77mp9.30.

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. Proc FATE. https://doi.org/10.1145/3287560.3287598

Shneiderman B (2020) Human-centered artificial intelligence: reliable, safe, and trustworthy. Int J Hum Comput Interact 36(6):495–504. https://doi.org/10.1080/10447318.2020.1741118

Solak MC (2025) The societal epistemology of artificial intelligence: the rise of algorithmic authority in the production of knowledge in the social sciences. İnsan Ve Toplum Bilimleri Araştırmaları Dergisi 14(4):144–160

Stevenson MT, Doleac JL (2022) Algorithmic risk assessment in the hands of humans. J Law Econ 65(4):735–770. https://doi.org/10.1086/720377

Suchman MC (1995) Managing legitimacy: strategic and institutional approaches. Acad Manag Rev 20(3):571–610. https://doi.org/10.5465/amr.1995.9508080331

Suchman L (2007) Human–machine reconfigurations: plans and situated actions, 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511808418

Suri K (2025) From decision support to decision substitution: a behavioural framework for AI overreach in leadership. SSRN. https://ssrn.com/abstract=6026314

Tronto JC (1993) Moral boundaries: a political argument for an ethic of care. Routledge. https://doi.org/10.4324/9780203730622

Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190498511.001.0001

Verbeek PP (2024) Moralizing technology: understanding and designing the morality of things. University of Chicago Press

Waggoner DB, Neely AD, Kennerley MP (1999) The forces that shape organisational performance measurement systems. Int J Prod Econ 60:53–60. https://doi.org/10.1016/S0925-5273(98)00201-1

Williamson B, Eynon R (2020) Historical threads, missing links, and future directions in AI in education. Learn Media Technol 45(3):223–235. https://doi.org/10.1080/17439884.2020.1798995

Yeung K (2018) Algorithmic regulation: a critical interrogation. Regul Gov 12(4):505–523. https://doi.org/10.1111/rego.12158

Zuboff S (2023) The age of surveillance capitalism. Social theory rewired. Routledge, pp 203–213. https://doi.org/10.4324/9781003320609

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.