E-ISSN: 3108-4192

APSSHS

Academic Publications of Social Sciences and Humanities Studies
2024, Volume 4, Issue 2, Page No: 121-129
Available online at: https://apsshs.com/

Journal of Applied Organizational Systems and Behavior

Human Judgment and Responsibility after Artificial Intelligence Recommendations

Yifan Zhao[1]*, Liang Chen[1], Ahmad Fauzi[2], Nor Hayati[3]

1. Department of Human Judgment and AI Responsibility, Faculty of Business, China Agricultural University, Beijing, China.
2. Department of AI Recommendations and Decision-Making, Faculty of Management, University of Malaya, Kuala Lumpur, Malaysia.
3. Department of Organizational Responsibility, Faculty of Psychology, Universiti Putra Malaysia, Serdang, Malaysia.

Abstract

Artificial intelligence recommendations increasingly enter organizational decisions that remain formally assigned to human actors. Yet prevailing accounts often treat reliance as a problem of accuracy, trust, adoption, or interface design without examining how machine advice may reshape who performs judgment, who remains psychologically connected to consequences, and who accepts responsibility afterward. This Original Propositional Model Article develops a proposed Judgment-and-Responsibility Model integrating decision deference, judgment displacement, moral distance, and responsibility avoidance. The model distinguishes ordinary advice use from deference, and deference from delegation, arguing that responsibility becomes vulnerable not whenever advice is accepted, but when independent appraisal and justificatory ownership are displaced. It further proposes that judgment displacement may increase moral distance by reducing the salience of affected persons, contextual particulars, or downstream consequences, thereby creating conditions in which responsibility can be avoided or diffusely assigned. These relationships are conditional on task characteristics, expertise, verification difficulty, organizational authority, consequence visibility, work design, and opportunities to contest or review recommendations. The article advances propositions and organizational safeguards centered on independent justification, traceability, consequence reconnection, explicit responsibility assignment, and meaningful human control. The synthesis is non-empirical and does not establish a validated causal model, universal prescription, or implementation-ready system. Its principal contribution is to connect scholarship on human–AI reliance with organizational responsibility while specifying rival explanations, multilevel boundaries, and empirical designs capable of refining, challenging, or rejecting the proposed relationships.

Keywords: Human judgment after machine advice, Decision deference, Deference to artificial intelligence recommendations, Moral distance from decision consequences, Theoretical propositions, Organizational safeguards and empirical tests

How to cite this article: Zhao Y, Chen L, Fauzi A, Hayati N. Human Judgment and Responsibility After Artificial Intelligence Recommendations. J Appl Organ Syst Behav. 2024;4(2):121-9. https://doi.org/10.51847/5UBPUjO0HH

Received: 28 May 2024; **Revised:** 25 August 2024; **Accepted:** 27 August 2024
Corresponding author: Yifan Zhao
E-mail ✉ yifan.zhao@cau.edu.cn

Introduction

Artificial intelligence recommendations increasingly participate in decisions about workers, customers, patients, applicants, risks, and resources. Such systems do more than supply information. They may structure attention, prioritize cases, classify conduct, and make some courses of action appear procedurally normal. Research on algorithms at work shows that these technologies can reorganize direction, evaluation, discipline, and contestation rather than merely automate isolated calculations [1]. The central organizational problem is therefore not simply whether a recommendation is accurate. It is

whether human actors continue to exercise judgment in a form that remains explainable, challengeable, and connected to responsibility when algorithmic advice becomes embedded in routine authority structures.

Trust is an important part of this problem, but it is not an adequate substitute for judgment. Empirical research treats trust in artificial intelligence as multidimensional and sensitive to characteristics of the system, task, and observed performance [2]. A person may trust a system yet inspect its recommendation carefully, or distrust it while still complying because of organizational pressure. Likewise, visible acceptance does not reveal whether the user independently evaluated the advice, deferred to perceived technical authority, or followed a formal rule. The article consequently separates trust, reliance, compliance, deference, and delegation so that responsibility is not inferred from behavior that may have several different explanations.

The management literature also warns against presenting automation and augmentation as mutually exclusive organizational choices. They are interdependent and may coexist within the same decision process, creating tensions between increased machine authority and continuing expectations of human accountability [3]. A nominally augmented decision can become substantively automated when the human contribution is limited to confirming an already privileged output. Conversely, automation of information processing can strengthen judgment when it preserves room for interpretation, contextual reasoning, and challenge. The relevant question is therefore not whether humans remain somewhere in the workflow, but what cognitive, moral, and organizational work they are still expected and enabled to perform.

Human–AI complementarity offers a constructive alternative to replacement narratives because machine analysis and human contextual judgment may contribute different capabilities under uncertainty [4]. Nevertheless, complementarity should not be assumed merely because both actors are present. It depends on how tasks are allocated, whether differences in capability are understood, and whether the human retains a meaningful obligation to justify the final decision. Organizational structures further shape this obligation by allocating decision rights, escalation authority, and accountability across people and systems [5]. Formal responsibility may remain attached to a role even when enacted authority has shifted toward an algorithmic recommendation.

This article develops an Original Propositional Model Article explaining how machine advice may influence judgment and responsibility. Its proposed contribution is the Judgment-and-Responsibility Model, which connects decision deference to judgment displacement, moral distance, and responsibility avoidance while retaining alternative pathways of active engagement and responsibility retention. The model does not presume that accepting AI advice is irresponsible or that delegation is inherently evasive. Instead, it specifies conditional relationships, rival explanations, level-of-analysis boundaries, and safeguards that require empirical testing. The aim is to clarify when AI recommendations remain evidence for human judgment and when organizational arrangements may allow them to become substitutes for judgment without a corresponding redistribution of legitimate responsibility.

Theoretical foundations of decision deference

Decision deference is proposed here as the degree to which an external recommendation determines a final judgment without equivalent independent appraisal. It differs from simple advice use because a recommendation can influence a decision while the user still evaluates its assumptions, compares alternatives, and owns the reasoning. Experimental evidence on algorithm appreciation indicates that people may assign greater weight to algorithmic estimates than to comparable human estimates under some conditions [6]. This finding establishes neither universal preference nor excessive reliance. It instead demonstrates that machine origin can affect advice weighting, making deference a plausible but context-dependent feature of judgment rather than an inevitable response to technical systems.

Perceived control is one condition that can alter reliance. Allowing users to modify an imperfect algorithm, even slightly, can increase willingness to use it [7]. This effect may reflect ownership, reduced reactance, or a belief that local knowledge can correct system limitations. It also shows why system use should not be interpreted as surrender of judgment: modification can represent active engagement rather than passive acceptance. At the same time, nominal control can become symbolic when adjustments are technically possible but organizationally discouraged, cognitively burdensome, or unlikely to affect the final decision. The proposed model therefore distinguishes formal permission to intervene from substantive capacity to revise, reject, or explain a recommendation.

Task perceptions provide another boundary. Algorithm aversion is stronger when tasks are perceived as subjective rather than objective [8]. This pattern may arise because subjective judgments appear to require empathy, contextual interpretation, or tacit knowledge that users do not attribute to machines. Yet perceived objectivity can also create misplaced confidence when measurement choices, categories, and values are concealed inside a numerical output. Task type should consequently be treated as an interpreted property shaped by occupational norms and interface framing, not as a fixed technical fact. Deference may be rational when relative capability is known, but apparent objectivity alone cannot establish that independent appraisal is unnecessary.

At the individual level, deference concerns how a person weighs and interrogates advice. At team and organizational levels, it concerns norms, decision rights, workload, and the consequences of disagreement. An employee may appear deferential because the system is trusted, because verification is difficult, because deviation requires authorization, or because acceptance offers procedural protection. The proposed construct therefore requires behavioral and process measures that separate recommendation acceptance from reduced reasoning, justification, and ownership. It also requires temporal analysis: repeated reliable advice may appropriately increase reliance, whereas routinized acceptance may persist after task conditions or system performance change. Decision deference is thus a proposed relational state between actor, recommendation, task, and organization rather than a stable personal disposition.

Deference to artificial intelligence recommendations

Deference to artificial intelligence recommendations can take the form of under-deference, calibrated reliance, or over-deference. Under-deference occurs when advice is discounted despite relevant capability; over-deference occurs when it is accepted without sufficient scrutiny; calibrated reliance requires that advice weight correspond to task-relevant evidence and known limitations. Resistance to medical artificial intelligence illustrates that under-deference may arise when users believe a system neglects individual uniqueness [9]. Such resistance is not necessarily irrational. It can express legitimate concern about contextual omission, although it may also persist when personalization is technically available. This evidence supports a contingent account in which perceptions of person-specific understanding influence whether machine advice is considered appropriate.

Over-deference becomes more plausible when verification is difficult. A systematic review of automation bias identifies verification complexity as an important condition affecting whether users detect and correct erroneous automated advice [10]. Complexity can increase cognitive burden, obscure the basis of a recommendation, and make acceptance less costly than independent reconstruction. However, verification burden should not be treated as the sole cause of automation bias. Training, interface design, workload, time pressure, and expectations about system reliability may also explain acceptance. The proposed model therefore treats verification complexity as one antecedent of judgment displacement, not as proof that reliance has become irresponsible.

Reliance can also be selective rather than uniformly high or low. Experimental evidence from public-sector decision scenarios shows that automation bias can coexist with selective adherence to advice that aligns with prior preferences [11]. This finding complicates accounts that attribute deference only to perceived machine superiority. A recommendation may be followed because it legitimizes a favored position, offers procedural cover, or reduces the political cost of a choice. Conversely, incongruent advice may be rejected despite equal technical quality. Selective adherence therefore links epistemic reliance to motivation and organizational incentives, while still leaving open the alternative explanation that decision-makers are rationally integrating information unavailable to the system.

The article proposes deference calibration as a way to analyze these patterns without assuming that more or less reliance is inherently better. Calibration concerns whether the degree and form of reliance are appropriate to known capability, uncertainty, stakes, and verification opportunities. It also concerns the quality of human engagement: asking what evidence was considered, what context may be missing, and why the recommendation should govern the case at hand. This proposed interpretation moves beyond visible acceptance toward the process by which a decision becomes attributable to a human actor. It remains uncertain whether the same mechanisms operate across occupations, low- and high-stakes tasks, individual and team decisions, or voluntary and mandatory systems.

Moral distance from decision consequences

Moral distance refers to reduced psychological, informational, or organizational connection between a decision-maker and those affected by a decision. It is not identical to physical separation. Distance can arise when consequences are represented through categories, scores, or workflow states that suppress personal and contextual particulars. Experimental research shows that people are more averse to machines making moral decisions than nonmoral decisions [12]. This reaction suggests that perceived moral authority has boundaries distinct from technical capability. Yet aversion alone does not demonstrate that human decisions are more ethical, nor does it reveal what happens when a machine recommendation is advisory while the human retains formal approval authority.

Conceptual work on artificial intelligence and the ethics of care distinguishes proximity distance from bureaucratic distance [13]. Proximity distance concerns reduced visibility of affected persons and their circumstances. Bureaucratic distance concerns procedural layering that makes consequences appear to result from a system, rule, or distributed process rather than a situated choice. These forms may reinforce one another when users encounter only abstract outputs and must process cases quickly. The proposed model treats moral distance as a possible consequence of judgment displacement, but this relationship remains unvalidated. AI mediation may sometimes reduce distance by revealing overlooked risks, disparities, or downstream effects.

Responsibility gaps further complicate interpretation. Philosophical analysis of tragic choices argues that technology-related responsibility gaps can have competing normative implications rather than being uniformly objectionable [14]. In some settings, refusing to assign full blame to a single actor may recognize limited control, unavoidable uncertainty, or genuinely distributed action. In others, the same gap may permit institutions to avoid explanation and remedy. The article therefore distinguishes justified limits on individual blame from organizational responsibility avoidance. Moral responsibility, causal contribution, role responsibility, answerability, and liability should not be collapsed, because each may be allocated differently after an AI-supported decision.

The proposed relationship between deference and moral distance is conditional on consequence salience. When affected persons, contextual particulars, and likely outcomes remain visible, reliance on AI advice need not diminish ethical attention. When these elements are hidden, fragmented, or treated as outside the decision-maker's role, over-deference may make consequences feel remote and authorship ambiguous. Alternative explanations include workload, bureaucratic hierarchy, professional socialization, and pre-existing organizational distance. Testing the proposed pathway therefore requires measures of contextual recall, perceived proximity, ownership, justification quality, and willingness to revisit a decision, rather than relying only on self-reported trust or recommendation acceptance.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for definitions, constructs, and conceptual boundaries for human judgment and responsibility after artificial intelligence recommendations.

Table 1. Definitions, constructs, and conceptual boundaries for Human Judgment and Responsibility After Artificial Intelligence Recommendations.

Construct or component	Definition	Problem addressed	Distinction from adjacent concepts	Proposed role	Uncertainty	Boundary statement
Human judgment after machine advice	Contextual evaluation after receiving AI input	Human involvement may become ceremonial	More than final approval	Core decision activity	Quality is difficult to observe	Human presence does not prove meaningful judgment
Trust	Willingness to accept vulnerability to AI	Confidence may be mistaken for responsibility transfer	Attitude rather than behavior	Antecedent of reliance	May coexist with scrutiny	Trust does not equal deference
Advice reliance	Incorporation of a recommendation into judgment	Acceptance reveals little about reasoning	Influence rather than surrender	Observable decision input	Appropriateness is task dependent	Reliance may be calibrated
Decision deference	Recommendation governs with reduced independent appraisal	Judgment may shift without formal delegation	Stronger than advice use	Proposed precursor	Requires process measures	Acceptance alone is insufficient evidence
Delegation	Transfer of a task or decision authority	Authority may move to an artificial agent	Prospective transfer rather than advice weighting	Proposed structural condition	Motive may be instrumental	Delegation is not automatically avoidance
Judgment displacement	Independent reasoning becomes secondary or post hoc	Formal accountability may remain after substantive judgment shifts	Distinct from reliance and delegation	Proposed linking mechanism	Not empirically validated	Must be measured directly
Moral distance	Reduced connection to affected persons or consequences	Abstraction may weaken ethical attention	Not merely physical remoteness	Proposed mediator	AI may also reduce distance	No established causal effect
Responsibility avoidance	Reduced ownership, explanation, or answerability	AI may become a proposed means of blame protection	Differs from justified limits on blame	Proposed outcome pathway	Motivation may be concealed	Must be separated from responsibility diffusion
Responsibility retention	Capacity and obligation to justify, challenge, and own a decision	Human oversight may remain nominal	More than human presence	Proposed alternative model state	Organizational enactment varies	Not guaranteed by formal policy

Responsibility avoidance

Responsibility avoidance is proposed as a motivated or structurally enabled reduction in willingness to own, explain, revisit, or bear the consequences of a decision. It should not be inferred from delegation alone. Experimental research indicates that perceived agency and decision context influence willingness to delegate choices to autonomous AI [15]. Delegation may reflect superior expected performance, effort conservation, or legitimate task allocation. The relevant responsibility question arises when the arrangement also permits the human actor to claim that the outcome belongs primarily to the system while retaining authority, benefit, or formal status.

Behavioral evidence shows that delegation to an artificial agent can reduce punishment directed at the human delegator after a harmful outcome [16]. This result identifies a possible external attribution advantage, not a general organizational rule.

Punishment judgments differ from legal liability, professional accountability, and an actor's own sense of responsibility. Nevertheless, anticipated blame reduction may make machine delegation attractive when a decision is morally contested or likely to be reviewed. The proposed model therefore treats blame insulation as one pathway through which responsibility avoidance may become strategically relevant.

Recent experiments further indicate that indirect delegation to machine agents can increase dishonest instructions and compliance by lowering the moral cost of requesting misconduct [17]. The finding is bounded to the tested behavioral settings and does not establish that AI generally makes organizational actors dishonest. It supports the narrower proposition that interface structure and indirect instruction can alter experienced authorship. Responsibility avoidance may accordingly be motivated, as when an actor seeks distance from blame, or structural, as when roles and workflows obscure who was expected to judge. Distinguishing these forms requires direct measures of motive, perceived authorship, explanation willingness, and anticipated consequences.

Responsibility avoidance also differs by level of analysis. An individual may seek personal insulation, while a team may normalize collective reliance and an organization may design accountability so diffusely that no actor can reconstruct the decision. These patterns should not be collapsed into a single motive. Formal responsibility can remain assigned to a manager even when information access, discretion, and review authority are distributed elsewhere. The proposed model therefore asks whether actors could reasonably know, challenge, and alter the recommendation, as well as whether they were expected to answer for the result.

Proposed judgment-and-responsibility model

Learning algorithms can reconfigure organizational expertise, knowledge, and control relationships rather than operating as neutral inputs [18]. The proposed model therefore begins with four sets of conditions: recommendation characteristics, task characteristics, human characteristics, and organizational characteristics. These conditions influence appraisal and produce under-deference, calibrated reliance, or over-deference. None is inherently responsible or irresponsible; evaluation depends on capability, uncertainty, stakes, and whether the user can interrogate the recommendation.

Professional responses to opaque AI demonstrate that engagement is enacted through practices of questioning, comparison, workaround, or disengagement rather than determined by opacity alone [19]. The model treats active engagement and independent justification as an alternative pathway toward responsibility retention.

Figure 1 presents the original conceptual synthesis for judgment-and-responsibility model, showing how the article's principal constructs, mechanisms, uncertainties, and decision boundaries are connected.

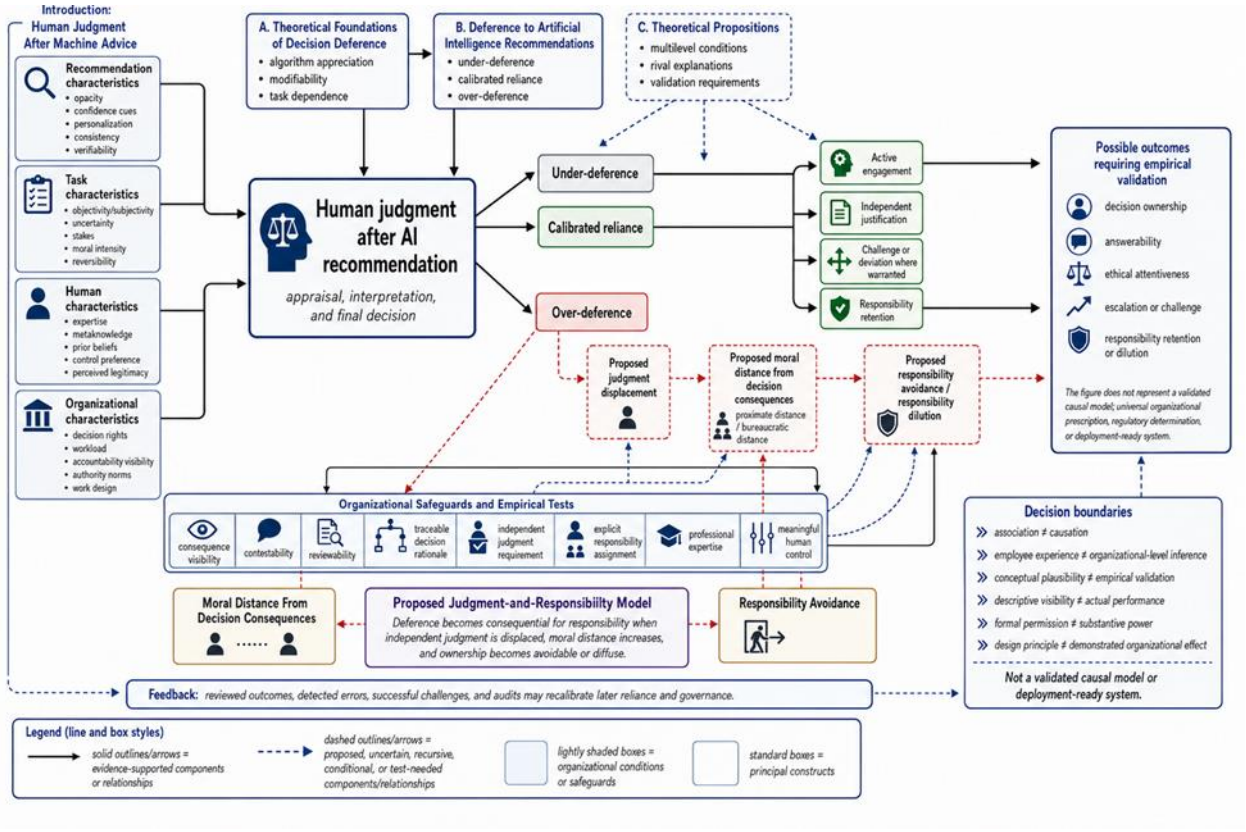

Figure 1. Judgment-and-Responsibility Model

Delegation to agentic information systems involves transfers of goals, authority, and action, making accountability an architectural rather than merely attitudinal question [20]. The model proposes judgment displacement when independent appraisal becomes secondary, ceremonial, or reconstructed after acceptance. This proposed mechanism connects over-deference to responsibility consequences without assuming that all reliance transfers judgment.

Algorithmic management can reshape autonomy, feedback, demands, and opportunities for intervention [21]. Work design may therefore strengthen or interrupt displacement by determining whether challenge is feasible and consequential.

The model also includes a recursive feedback process. Reviewed outcomes, detected errors, successful challenges, and organizational audits may recalibrate later reliance, revise decision rights, or expose ceremonial oversight. Apparent success may instead reinforce deference even when outcome quality is difficult to observe. Feedback must therefore be interpreted cautiously: repeated acceptance can reflect learning, but it can also reflect declining scrutiny, institutional dependence, or selection of cases that conceal system limitations.

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for proposed mechanisms, relationships, and organizational consequences in human judgment and responsibility after artificial intelligence recommendations.

Table 2. Proposed mechanisms, relationships, and organizational consequences in Human Judgment and Responsibility After Artificial Intelligence Recommendations.

Mechanism or relationship	Antecedent	Mediator or process	Moderator	Expected consequence	Rival explanation	Validation requirement
Algorithmic authority to advice weight	Perceived competence	Reliance	Expertise and task	Greater recommendation influence	Rational capability use	Source-randomized experiment
Control to adoption	Modification opportunity	Ownership or reduced reactance	Meaningful discretion	Increased use	Interface preference	Compare symbolic and effective control
Verification burden to over-deference	Complexity	Reduced checking	Workload and expertise	Missed error	Time pressure	Process tracing
Preference alignment to selective adherence	Congruent advice	Motivated acceptance	Incentives and prior belief	Directional reliance	Rational updating	Hold evidence quality constant
Judgment displacement to moral distance	Reduced appraisal	Lower consequence salience	Visibility and moral intensity	Weaker felt authorship	Pre-existing bureaucracy	Mediation experiment
Moral distance to responsibility avoidance	Abstracted consequences	Blame insulation	Role clarity and review	Reduced ownership	Performance optimization	Measure motive and attribution
Contestability to responsibility retention	Challenge opportunity	Independent justification	Authority and psychological safety	Greater answerability	Documentation compliance	Longitudinal field test

Taken together, organizational embedding, delegation architecture, and work design make the proposed pathway multilevel [18, 20, 21]. Solid model components are evidence-supported; judgment displacement and its cross-domain links remain proposed. The model can be rejected if deference does not reduce independent appraisal, if moral distance does not mediate responsibility outcomes, or if observed effects are fully explained by rational capability allocation, workload, or formal authority.

Theoretical propositions

Hybrid intelligence locates value in complementary human and machine capabilities rather than simple substitution [22]. **Proposition 1** states that perceived AI competence will increase deference, but the relationship will be weaker when task subjectivity, contextual uniqueness, or meaningful modification rights make independent judgment more salient. **Proposition 2** states that verification complexity will strengthen the relationship between advice acceptance and proposed judgment displacement, conditional on workload and expertise.

Treating machines as teammates introduces coordination, role, and shared-responsibility questions beyond individual advice weighting [23]. **Proposition 3** states that team norms portraying AI acceptance as procedurally protective will strengthen responsibility avoidance. **Proposition 4** states that explicit decision ownership and consequential dissent rights will weaken this relationship. These are multilevel propositions: team norms should not be inferred from individual attitudes, and formal permission to challenge does not establish substantive influence.

Experimental evidence indicates that productive delegation depends on metaknowledge about relative human and AI capabilities [24]. **Proposition 5** states that metaknowledge will increase calibrated delegation while reducing both under-deference and over-deference. **Proposition 6** states that capability-based delegation will be less associated with responsibility avoidance when the human remains required to justify task allocation and outcome review.

The combined logic is that complementarity, team coordination, and metaknowledge condition productive human–AI judgment [22–24]. **Proposition 7** states that judgment displacement will increase moral distance only when consequence salience is low. **Proposition 8** states that moral distance will increase responsibility avoidance only when technical or organizational arrangements provide plausible blame insulation. **Proposition 9** states that contestability, reviewability, and traceable rationale will interrupt this pathway. Each proposition remains open to null, reversed, occupation-specific, or nonlinear relationships.

The propositions also require discriminant tests. Judgment displacement should predict reduced independent reasoning beyond trust, advice weight, time pressure, and perceived accuracy. Moral distance should be distinguishable from low empathy, physical remoteness, and routine bureaucratic detachment. Responsibility avoidance should predict reduced willingness to explain, revisit, or accept consequences beyond ordinary delegation and legitimate role limits. Failure to establish these distinctions would require narrowing or abandoning the proposed constructs rather than relabelling familiar reliance effects.

Organizational safeguards and empirical tests

Human-centered AI scholarship argues that high automation can coexist with high human control [25]. The proposed safeguards therefore do not require rejecting AI recommendations. They require preserving meaningful judgment through independent justification, accessible uncertainty, override authority, traceable acceptance and deviation, and explicit responsibility assignment. Consequence-reconnection practices should make affected persons and likely outcomes visible at the point of decision.

An integrated view of algorithm aversion and appreciation indicates that reliance depends on interacting user, task, system, and contextual conditions [26]. Empirical tests should consequently measure advice weight, verification behavior, reasoning before and after advice, justification quality, perceived authorship, consequence salience, and willingness to revisit outcomes. Randomized experiments can test source, accuracy, opacity, and responsibility allocation; longitudinal field studies can examine habituation, challenge practices, and recalibration.

Responsible-AI governance encompasses structural, relational, and procedural practices [27]. Comparative organizational studies should test whether formal controls are enacted, whether dissent changes decisions, and whether responsibility remains traceable across lifecycle stages. Multilevel designs are needed to separate individual reliance from team norms and organizational authority.

The proposed safeguard configuration combines meaningful human control, context-sensitive reliance calibration, and multidimensional governance [25–27]. It is not an implementation prescription. Effects may vary by occupation, stakes, regulation, system maturity, and institutional power. Evidence should be capable of refining or rejecting safeguards, including findings that documentation becomes ceremonial, overrides are punished, or consequence visibility increases defensive rather than responsible judgment.

Temporal ordering is equally important. Studies should measure appraisal before advice, reasoning during exposure, acceptance or rejection, subsequent justification, and later responsibility attribution. Cross-sectional attitudes cannot establish whether deference preceded moral distance or whether actors retrospectively distanced themselves after an unfavorable outcome. Repeated-measures designs should test whether reliable performance produces calibrated learning or habitual acceptance, and whether errors restore scrutiny. Pre-registered rival models should allow workload, authority pressure, expected accuracy, and strategic self-protection to compete with the proposed pathway.

Safeguards can themselves fail. Independent-judgment forms may invite post hoc rationalization; traceability may produce defensive documentation; override rights may remain unused when dissent threatens status or workload; and responsibility assignment may concentrate blame on users who lack real authority. Evaluation should therefore compare formal safeguards with enacted practice. Field experiments, interrupted time-series designs, qualitative process studies, and audit-log analyses can examine whether safeguards change verification, challenge, and ownership rather than merely increasing visible compliance.

Conclusion

This article proposes the Judgment-and-Responsibility Model to explain why the organizational significance of AI recommendations extends beyond accuracy, trust, and adoption. Its central distinction is between using machine advice and displacing human judgment. The model proposes that over-deference may affect responsibility when independent appraisal

is weakened, consequences become morally distant, and technical or organizational arrangements permit ownership to be avoided or diffused.

These relationships remain conditional and unvalidated. Delegation may be rational, responsibility may be genuinely distributed, and AI may sometimes improve consequence visibility or human judgment. The model therefore preserves alternative pathways of calibrated reliance, active engagement, independent justification, and responsibility retention. Its propositions and safeguards are intended as falsifiable conceptual claims rather than universal prescriptions. The contribution lies in connecting epistemic reliance to moral and organizational responsibility while identifying the task, human, team, work-design, and governance conditions that must be measured before causal or practical conclusions can be justified. Accordingly, retaining a human approver is not sufficient; responsibility depends on whether judgment remains practicable, consequential, and owned.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Kellogg KC, Valentine MA, Christin A. Algorithms at work: The new contested terrain of control. Acad Manag Ann. 2020;14(1):366-410. doi:10.5465/annals.2018.0174
2. Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. Acad Manag Ann. 2020;14(2):627-60. doi:10.5465/annals.2018.0057
3. Raisch S, Krakowski S. Artificial intelligence and management: The automation–augmentation paradox. Acad Manage Rev. 2021;46(1):192-210. doi:10.5465/amr.2018.0072
4. Jarrahi MH. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Bus Horiz. 2018;61(4):577-86. doi:10.1016/j.bushor.2018.03.007
5. Shrestha YR, Ben-Menahem SM, von Krogh G. Organizational decision-making structures in the age of artificial intelligence. Calif Manage Rev. 2019;61(4):66-83. doi:10.1177/0008125619862257
6. Logg JM, Minson JA, Moore DA. Algorithm appreciation: People prefer algorithmic to human judgment. Organ Behav Hum Decis Process. 2019;151:90-103. doi:10.1016/j.obhdp.2018.12.005
7. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: People will use imperfect algorithms if they can—even slightly—modify them. Manage Sci. 2018;64(3):1155-70. doi:10.1287/mnsc.2016.2643
8. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. J Mark Res. 2019;56(5):809-25. doi:10.1177/0022243719851788
9. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res. 2019;46(4):629-50. doi:10.1093/jcr/ucz013
10. Lyell D, Coiera E. Automation bias and verification complexity: A systematic review. J Am Med Inform Assoc. 2017;24(2):423-31. doi:10.1093/jamia/ocw105
11. Alon-Barkat S, Busuioc M. Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. J Public Adm Res Theory. 2023;33(1):153-69. doi:10.1093/jopart/muac007
12. Bigman YE, Gray K. People are averse to machines making moral decisions. Cognition. 2018;181:21-34. doi:10.1016/j.cognition.2018.08.003
13. Villegas-Galaviz C, Martin K. Moral distance, AI, and the ethics of care. AI Soc. 2024;39(4):1695-706. doi:10.1007/s00146-023-01642-z
14. Danaher J. Tragic choices and the virtue of techno-responsibility gaps. Philos Technol. 2022;35(2):26. doi:10.1007/s13347-022-00519-1
15. Candrian C, Scherer A. Rise of the machines: Delegating decisions to autonomous AI. Comput Hum Behav. 2022;134:107308. doi:10.1016/j.chb.2022.107308
16. Feier T, Gogoll J, Uhl M. Hiding behind machines: Artificial agents may help to evade punishment. Sci Eng Ethics. 2022;28(2):19. doi:10.1007/s11948-022-00372-7
17. Köbis N, Rahwan Z, Rilla R, Supriyatno BI, Bersch C, Ajaj T, et al. Delegation to artificial intelligence can increase dishonest behaviour. Nature. 2025;646(8083):126-34. doi:10.1038/s41586-025-09505-x

18. Faraj S, Pachidi S, Sayegh K. Working and organizing in the age of the learning algorithm. Inf Organ. 2018;28(1):62-70. doi:10.1016/j.infoandorg.2018.02.005
19. Lebovitz S, Lifshitz-Assaf H, Levina N. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. Organ Sci. 2022;33(1):126-48. doi:10.1287/orsc.2021.1549
20. Baird A, Maruping LM. The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. MIS Q. 2021;45(1):315-41. doi:10.25300/MISQ/2021/15882
21. Parent-Rocheleau X, Parker SK. Algorithms as work designers: How algorithmic management influences the design of jobs. Hum Resour Manag Rev. 2022;32(3):100838. doi:10.1016/j.hrmr.2021.100838
22. Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid intelligence. Bus Inf Syst Eng. 2019;61(5):637-43. doi:10.1007/s12599-019-00595-2
23. Maier R, Seeber I, Bittner E, Briggs RO, de Vreede T, de Vreede GJ, et al. Machines as teammates: A research agenda on AI in team collaboration. Inf Manag. 2020;57(2):103174. doi:10.1016/j.im.2019.103174
24. Fügener A, Grahl J, Gupta A, Ketter W. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Inf Syst Res. 2022;33(2):678-96. doi:10.1287/isre.2021.1079
25. Shneiderman B. Human-centered artificial intelligence: Reliable, safe and trustworthy. Int J Hum Comput Interact. 2020;36(6):495-504. doi:10.1080/10447318.2020.1741118
26. Jussupow E, Benbasat I, Heinzl A. An integrative perspective on algorithm aversion and appreciation in decision-making. MIS Q. 2024;48(4):1575-90. doi:10.25300/MISQ/2024/18512
27. Papagiannidis E, Mikalef P, Conboy K. Responsible artificial intelligence governance: A review and research framework. J Strateg Inf Syst. 2025;34(2):101885. doi:10.1016/j.jsis.2024.101885