Human Judgment in A.I.-Assisted Management Decisions

Thomas A. Guerriero, PhD

Working Paper | 21 November 2026

Abstract

Managers in nonprofit and non-governmental organisations increasingly receive advice from artificial intelligence tools when they forecast demand, allocate staff and budgets, screen applications, prioritise programmes and assess risk. The central management question is no longer whether such advice is available but how human judgement should be combined with it. This conceptual paper develops a calibrated-judgement framework for AI-assisted management decisions. It builds on two experimental literatures that point in different directions. Dietvorst, Simmons and Massey report that people became less likely to rely on an algorithm after seeing it err, even when it outperformed a human forecaster. Alon-Barkat and Busuioc, in three experiments in the Netherlands, found no evidence of automation bias but found stronger adherence to advice aligned with group stereotypes in one study, with no significant difference between algorithmic and human-expert advice. These findings are connected to research on accountability and advice rationales, conditions for intuitive expertise, complacency and automation bias, decision structures and the automation–augmentation paradox. The framework holds that reliance on AI advice should be calibrated to the predictability of the decision environment, the reasons given for advice and the accountability that attaches to the decision. Six propositions address advice rationales, track records, selective adherence, workload, decision structure and skill preservation. No original empirical data are presented.

Keywords: human judgement; algorithmic advice; managerial decision making; artificial intelligence; nonprofit management

Introduction

Artificial intelligence tools now offer advice on many of the decisions that managers make. A tool may forecast how many people will seek a service next month, recommend how to allocate volunteers across sites, score applications for a programme, flag a project report for closer review or estimate the risk that a project will run late. In nonprofit and non-governmental organisations, where managers often handle many responsibilities with limited support, such advice can be welcome. It can also be wrong, biased or poorly suited to the case at hand.

The management question is how human judgement should be combined with AI advice. Two simple answers are tempting and both are inadequate. The first is that managers should follow the tool because algorithms are more consistent than people. The second is that managers should treat the tool as a suggestion they may ignore at will. The first risks automating errors and biases. The second risks discarding useful information because of distrust or habit. A better answer requires understanding how people actually respond to algorithmic advice and under what conditions human judgement adds value.

This paper draws on two experimental studies that illuminate how people respond to algorithmic advice. Dietvorst, Simmons and Massey (2015) report five studies in which participants who saw an algorithm make forecasts became less confident in it and less likely to choose it over an inferior human forecaster, even when they had seen the algorithm outperform the human. They describe this phenomenon as algorithm aversion. Alon-Barkat and Busuioc (2023) conducted three experiments in the Netherlands to test two potential biases in public sector decision making: automation bias, meaning overreliance on algorithmic advice despite warning signals, and selective adherence, meaning adoption of advice when it corresponds to stereotypes. They did not find evidence of automation bias. In their second study they found a stronger propensity to adhere to advice aligned with group stereotypes, with no significant difference between algorithmic and human-expert advice, but in a third study with civil servants, conducted shortly after a major scandal involving an algorithm with discriminatory outcomes, they did not find selective adherence.

Read together, these studies suggest that the problem is not simply that people trust algorithms too much or too little. Reliance can shift with experience of errors, with the fit between advice and prior beliefs and with the wider context. A framework for AI-assisted management decisions must therefore attend to the conditions under which reliance is appropriate.

An earlier working paper in this series considered AI in decisions about access to essential support in nonprofit human services, arguing for preserving human discretion where decisions must be explained, adapted and corrected. The present paper addresses a different domain: management decisions about the organisation itself, such as forecasting, allocation, prioritisation and risk assessment. These decisions also affect people, but they are typically internal, repeated

and subject to review by boards and stakeholders rather than by the individuals concerned. The paper's focus is therefore on calibration, meaning the match between how much a manager relies on advice and how reliable that advice is likely to be.

The paper makes three contributions. First, it brings together experimental evidence on algorithm aversion, automation bias and selective adherence with research on managerial accountability and expertise. Second, it proposes a calibrated-judgement framework that links reliance on AI advice to the decision environment, the reasons given for advice and accountability. Third, it develops six propositions suitable for empirical testing.

The analysis relies on verified references selected for relevance rather than a systematic review. Alon-Barkat and Busuioc (2023) and Dietvorst et al. (2015) were read in full from openly available copies. Claims about other cited works are limited to their verified abstracts, bibliographic records or, in the case of the NIST framework, the published report. The experimental findings discussed were obtained in specific settings, often with non-managerial participants, and are not treated as direct evidence about managers in nonprofit organisations. Where this paper uses the language of influence or effect, it refers to proposed relationships to be tested.

Literature Review

Algorithm aversion

Dietvorst et al. (2015) begin from evidence that evidence-based algorithms often predict more accurately than human forecasters, yet people frequently prefer human judgement when choosing between them. In five studies, participants saw an algorithm make forecasts, saw a human make forecasts, saw both or saw neither, and then decided whether to tie their incentives to the algorithm or to the human. In several studies the task involved predicting how well past applicants had performed in a master's of business administration programme. The authors report that participants who saw the algorithm perform were less confident in it and less likely to choose it over an inferior human forecaster, and that this was true even among participants who saw the algorithm outperform the human. They attribute the effect to people losing confidence in algorithmic forecasters more quickly than in human forecasters after seeing them make the same mistake.

The implication for management is that exposure to an algorithm's errors can lead to under-use of a tool that, on balance, performs well. A manager who notices a few memorable mistakes may abandon a tool whose overall record is better than the alternatives. This is not necessarily irrational in every case, since some errors may reveal real weaknesses, but it suggests that how errors are presented and how performance is tracked can matter.

Automation bias and selective adherence

Alon-Barkat and Busuioc (2023) investigate two biases drawn from psychology and public administration. Automation bias refers to overreliance on algorithmic advice even in the face of warning signals from other sources. Selective adherence refers to selective adoption of algorithmic advice when it corresponds to stereotypes. In study 1 (N = 605), they compared adherence to an algorithmic prediction with adherence to an equivalent human-expert prediction and did not find evidence of automation bias. In study 2 (N = 904), they replicated this result and found a stronger propensity for adherence when advice was aligned with group stereotypes, with no significant differences between algorithmic and human-expert advice. In study 3 (N = 1,345), conducted with civil servants shortly after the childcare benefits scandal, they supported their earlier finding on automation bias but did not find selective adherence, which they suggest may reflect heightened awareness of discrimination and algorithmic bias after the scandal. In study 3, participants were less likely to follow the algorithmic advice than the human-expert advice.

Two lessons follow for management. First, the assumption that people automatically defer to algorithms is not supported by these experiments. Second, the risk may lie less in deference to algorithms as such than in selective use of advice that confirms what decision-makers already believe. The third study also suggests that context and awareness can change patterns of reliance.

Complacency and automation bias under workload

Parasuraman and Manzey (2010) review empirical studies of complacency and bias in human interaction with automated and decision support systems. Their abstract reports that automation complacency occurs under conditions of multiple-task load, when manual tasks compete with the automated task for attention, and that it is found in both naive and expert participants and cannot be overcome with simple practice. Automation bias, in their account, results in both omission and commission errors when decision aids are imperfect, occurs in naive and expert participants, cannot be prevented by training or instructions, and can affect teams as well as individuals. They argue that both depend on similar attentional processes.

This review sits in some tension with Alon-Barkat and Busuioc’s (2023) failure to find automation bias. One possible reconciliation is that conditions matter. Experimental tasks in which participants consider a single case may not reproduce the multiple-task load that Parasuraman and Manzey identify as a condition for complacency. Managers in small organisations, who juggle many responsibilities, may be closer to the conditions under which complacency arises.

Accountability and advice rationales

Kotzian, Bauch and Weißenberger (2025) examine how accountability affects advice-taking from human and algorithmic advisers. According to their abstract, they manipulated the type of adviser, the presence of an advice rationale and the level of accountability, meaning whether

managers had to justify their judgements to superiors. For human advisers, receiving an advice rationale increased advice-taking more under high accountability than under low accountability. For algorithmic advisers, the positive effect of an advice rationale on advice-taking was not moderated by accountability but was consistently strong. The authors conclude that providing an advice rationale is essential for algorithmic advisers regardless of accountability.

For nonprofit managers, who are accountable to boards, partners, regulators and the people they serve, this finding is directly relevant. It suggests that tools that explain how they reached a recommendation may be used more readily. Whether they are used more appropriately, that is, followed when they are right and questioned when they are wrong, is a separate question that the abstract does not address.

Conditions for intuitive expertise

Kahneman and Klein (2009) explore differences between the heuristics and biases approach and naturalistic decision making. Their abstract states that evaluating the likely quality of an intuitive judgement requires an assessment of the predictability of the environment in which the judgement is made and of the individual's opportunity to learn the regularities of that environment, and that subjective experience is not a reliable indicator of judgement accuracy. This offers a principled basis for deciding when human judgement is likely to add value. Where the environment is predictable and managers have had the opportunity to learn its patterns, their judgement may be valuable. Where the environment is irregular or feedback is scarce, confidence may be a poor guide to accuracy, whether the judgement is the manager's or the manager's assessment of a tool.

Decision structures and the automation–augmentation paradox

Shrestha, Ben-Menahem and von Krogh (2019) compare human and AI-based decision making along five contingency factors: the specificity of the decision search space, the interpretability of the decision process and outcome, the size of the alternative set, decision-making speed and replicability. They outline three structural categories for combining human and AI decisions: full human-to-AI delegation, hybrid sequential decision making in which humans and AI pass decisions to one another, and aggregated human–AI decision making. Their framework is useful because it treats the combination of human and AI judgement as a design choice with several options rather than a binary between human and machine.

Raisch and Krakowski (2021) argue that, in the management domain, augmentation, in which humans collaborate closely with machines, cannot be neatly separated from automation, in which machines take over a task. They describe these as interdependent across time and space, creating a paradoxical tension, and argue that overemphasising either fuels reinforcing cycles with negative organisational and societal outcomes, while adopting a broader perspective that includes

both could achieve complementarities. Their argument suggests that the balance between human and AI roles should be revisited over time rather than fixed once.

Supporting human decision-makers

Anastasopoulos (2026), writing about public administration, argues according to the publisher's abstract that transferring bureaucratic discretion to algorithms threatens accountability, fairness and equity, individual consideration and reason-giving, and proposes a centaur model in which AI supports rather than replaces human officials. The NIST AI Risk Management Framework (Tabassi, 2023) offers a voluntary structure for managing AI risks organised around four functions: govern, map, measure and manage. Both sources support the view that human oversight should be designed deliberately, although neither addresses nonprofit management decisions specifically.

Conceptual Framework and Propositions

Calibrated judgement

The calibrated-judgement framework proposes that the appropriate combination of human judgement and AI advice in a management decision depends on three conditions.

Environmental predictability. Following Kahneman and Klein (2009), the framework asks whether the decision environment is regular enough, and feedback available enough, for either human intuition or algorithmic patterns to be reliable. Forecasting demand for a long-established service may be relatively predictable. Judging how a new programme will be received may not be.

Reason-giving. Following Kotzian et al. (2025) and Anastasopoulos (2026), the framework asks whether the advice comes with reasons that a manager can examine and, if necessary, explain to others. Advice without reasons is harder to calibrate, because the manager cannot assess whether the tool has considered the relevant factors.

Accountability. Following Kotzian et al. (2025), the framework asks to whom the manager must justify the decision and what form that justification takes. Accountability shapes how advice is used and creates a need for records that show how human judgement and AI advice were combined.

The framework defines **calibration** as the correspondence between how much a manager relies on AI advice and how reliable that advice is for the decision at hand. Over-reliance and under-reliance are both forms of miscalibration. The evidence reviewed above suggests that miscalibration can arise from several sources: memorable errors that drive under-use (Dietvorst et al., 2015), selective use of advice that confirms prior beliefs (Alon-Barkat & Busuioc, 2023), and complacency under workload (Parasuraman & Manzey, 2010).

Decision structure as a design choice

Following Shrestha et al. (2019), the framework treats the structure of human–AI decision making as a choice. Full delegation may suit decisions with a narrow search space, high replicability and low consequences. Hybrid sequential arrangements, in which the tool screens or ranks options and a manager decides, or in which a manager sets parameters and the tool executes, may suit many management decisions. Aggregated arrangements, in which human and AI judgements are combined, may suit decisions where both have relevant but different information. Following Raisch and Krakowski (2021), the framework holds that these choices should be revisited as tools, tasks and staff change.

Proposition 1: Advice rationales and appropriate reliance

Kotzian et al. (2025) report that advice rationales consistently increased advice-taking from algorithms. Increased advice-taking is beneficial only when advice is good. Rationales may also allow managers to detect when advice rests on irrelevant or mistaken factors. The framework proposes that rationales improve calibration, not merely reliance.

Proposition 1. Managers receiving AI advice accompanied by a rationale will show closer correspondence between their reliance on that advice and its accuracy than managers receiving the same advice without a rationale.

Proposition 2: Track records and recovery from errors

Dietvorst et al. (2015) report that seeing an algorithm err reduced confidence in it and reduced its selection, even when it outperformed a human. One response is to give managers a structured record of performance over time, comparing the tool with the human alternative, so that memorable errors are seen in context.

Proposition 2. Managers who receive periodic comparative records of AI and human performance on recurring decisions will be less likely to abandon a better-performing tool after observing its errors than managers who observe errors without such records.

Proposition 3: Selective adherence and independent first judgements

Alon-Barkat and Busuioc (2023) found, in one study, stronger adherence to advice aligned with group stereotypes, with no significant difference between algorithmic and human-expert advice. If managers adopt AI advice selectively when it confirms their prior views, the tool may reinforce rather than correct bias. Recording an independent judgement before seeing the advice may make selective adherence visible.

Proposition 3. In recurring management decisions, procedures that require managers to record an independent judgement before viewing AI advice will reduce the extent to which adherence to advice depends on its consistency with prior beliefs.

Proposition 4: Workload and complacency

Parasuraman and Manzey (2010) associate automation complacency with multiple-task load. Managers in small nonprofit organisations often carry many responsibilities at once. Under heavy workload, they may accept AI advice with less scrutiny, regardless of training.

Proposition 4. Managers' scrutiny of AI advice, measured by the frequency with which they detect deliberately introduced errors, will decline as concurrent workload increases, and this decline will not be eliminated by training alone.

Proposition 5: Matching decision structure to predictability

Kahneman and Klein (2009) link the quality of intuitive judgement to environmental predictability and the opportunity to learn. Shrestha et al. (2019) identify interpretability and specificity of the search space among the factors that shape how human and AI decisions should be combined. The framework proposes that organisations achieve better outcomes when they match decision structure to these conditions.

Proposition 5. Organisations that assign greater decision authority to AI tools in predictable, well-specified decisions, and greater authority to human judgement in irregular, poorly specified decisions, will achieve better decision outcomes than organisations that apply a single decision structure across both kinds of decision.

Proposition 6: Preserving judgement over time

Raisch and Krakowski (2021) warn that overemphasising automation can fuel reinforcing cycles with negative outcomes. One such cycle may be the erosion of managerial skill. If managers rely on tools for decisions they once made themselves, they may lose the experience needed to recognise when a tool is wrong. Periodically making decisions without the tool, or reviewing cases where human and AI judgements diverged, may preserve that capacity.

Proposition 6. Managers who periodically make or review recurring decisions without AI advice will retain greater ability to detect erroneous AI advice over time than managers who rely on AI advice continuously.

Discussion

Beyond trust and distrust

Public discussion of AI often frames the issue as trust: whether people trust AI too much or too little. The evidence reviewed here suggests that this framing is incomplete. Dietvorst et al. (2015) show that reliance can collapse after visible errors even when a tool performs well. Alon-Barkat and Busuioc (2023) show that, in their experiments, the more pressing risk was selective adherence rather than general deference, and that context could change this pattern. Parasuraman and Manzey (2010) show that complacency can arise from workload rather than from trust as

such. Calibration offers a more useful aim than trust because it asks whether reliance matches reliability in the specific decision.

Accountability as a resource

Accountability is often seen as a burden on nonprofit managers. The framework suggests it can also support better judgement. Kotzian et al. (2025) report that accountability shaped how advice rationales affected advice-taking from human advisers. More broadly, the need to explain decisions to boards and stakeholders can prompt managers to examine advice rather than accept it passively. Organisations can use this by asking managers to record, for significant decisions, what the tool advised, what they decided and why. Such records serve accountability and also create the comparative performance data that Proposition 2 envisages.

The limits of experimental evidence

Much of the evidence comes from experiments. Dietvorst et al. (2015) used forecasting tasks with incentivised participants, and Alon-Barkat and Busuioc (2023) used experimental vignettes with the public and civil servants. Experiments isolate mechanisms but may not reproduce the pressures, relationships and stakes of real management decisions. Managers may know the tools well, may discuss advice with colleagues and may face consequences that experiments cannot simulate. The propositions are therefore framed for testing in field settings as well as experiments.

Nonprofit specificity

The framework addresses nonprofit and non-governmental organisations, but most of the evidence comes from other settings. Nonprofit managers may differ in several ways. They often work with limited analytical support, may use tools provided by third parties without the ability to inspect them, and may face accountability to multiple stakeholders with different expectations. These differences could make calibration harder. They could also make human judgement more valuable, because managers in small organisations often hold contextual knowledge that tools lack. Whether nonprofit settings differ systematically in patterns of reliance is an open question.

Ethical stakes in management decisions

Management decisions are sometimes treated as lower-stakes than decisions about individuals. Yet decisions about where to allocate services, which applications to prioritise or which programmes to continue affect people. Alon-Barkat and Busuioc’s (2023) evidence on selective adherence to stereotype-consistent advice is a reminder that bias can enter through the interaction between human and tool, not only through the tool itself. Organisations should monitor outcomes of AI-assisted management decisions for unintended disparities, using approaches such as those described in the NIST AI Risk Management Framework (Tabassi, 2023).

Revisiting the division of labour

Raisch and Krakowski (2021) argue that automation and augmentation are interdependent over time. A tool introduced to support a manager may gradually take over more of the decision as confidence grows, and a manager may gradually lose the skills needed to oversee it. Conversely, a tool may be sidelined after a memorable error even when it remains useful. The framework recommends that organisations revisit the division of labour periodically, using records of performance and outcomes, rather than treating the initial design as permanent.

Practical Implications for Nonprofit and NGO Leaders

The framework suggests several starting points for leaders. These are design suggestions consistent with the argument, not evidence-based prescriptions.

First, list the management decisions in which AI advice is used or proposed. For each, leaders can ask how predictable the decision environment is, whether the tool explains its advice and to whom the decision must be justified.

Second, prefer tools that give reasons. Where a choice exists, tools that explain the basis for their recommendations may support both appropriate reliance and accountability.

Third, record decisions and advice together. For significant or recurring decisions, a brief record of what the tool advised, what the manager decided and why can support accountability and build a track record for assessing calibration.

Fourth, review performance comparatively. Periodically comparing the accuracy of AI advice with that of human judgement on recurring decisions can help managers see errors in context and avoid abandoning useful tools or relying on poor ones.

Fifth, use independent first judgements for sensitive decisions. Where decisions could be affected by stereotypes or prior beliefs, asking managers to record their own judgement before viewing AI advice can make selective adherence visible.

Sixth, attend to workload. Expecting overloaded managers to scrutinise every recommendation may be unrealistic. Leaders can identify decisions where scrutiny matters most and protect time for it.

Seventh, preserve judgement. Periodically making or reviewing decisions without the tool, and examining cases where human and AI judgements diverged, can help managers retain the skills needed to oversee AI advice.

Eighth, revisit the arrangement. As tools, tasks and staff change, the appropriate division of labour may change too. An annual review of AI-assisted decisions, reported to the board, can keep the arrangement deliberate.

Limitations and Future Research

This paper is conceptual and presents no original empirical data. Its propositions are theoretical and require testing.

The evidence base has limits. The two focal studies were experiments conducted with specific populations and tasks, and their findings may not transfer directly to nonprofit managers making real decisions. Claims about Kotzian et al. (2025), Kahneman and Klein (2009), Parasuraman and Manzey (2010), Shrestha et al. (2019), Raisch and Krakowski (2021) and Anastasopoulos (2026) are limited to their verified abstracts, because full texts could not be retrieved during preparation.

The framework simplifies. Environmental predictability, reason-giving and accountability are not the only conditions that matter. The quality of the tool, the data on which it was trained, the organisation's culture and the relationships among decision-makers are also likely to affect how human judgement and AI advice are combined. The framework isolates three conditions because they can be assessed and influenced by leaders.

The concept of calibration assumes that the reliability of advice can be known. In many management decisions, outcomes are delayed, ambiguous or influenced by many factors, making it difficult to determine whether advice was good. This limits the feasibility of some recommendations, such as comparative performance records, for decisions with poor feedback.

Future research could test the propositions in several ways. Field experiments in nonprofit organisations could compare decisions made with and without advice rationales and with and without independent first judgements. Longitudinal studies could examine whether comparative performance records reduce abandonment of useful tools after errors and whether periodic decisions without tools preserve managers' ability to detect errors. Research could also examine whether workload moderates scrutiny of AI advice among nonprofit managers and whether matching decision structure to predictability improves outcomes.

Conclusion

AI tools are becoming a routine source of advice for managers in nonprofit and non-governmental organisations. The question is how human judgement should be combined with that advice. This paper has argued that the aim should be calibration: reliance that matches the reliability of advice for the decision at hand. Drawing on experimental evidence on algorithm aversion, automation bias and selective adherence, and on research on accountability, expertise, decision structures and the automation–augmentation paradox, it has proposed a calibrated-judgement framework and six testable propositions. The central claim is that good AI-assisted management depends less on trusting or distrusting tools than on designing the conditions under which managers can judge when to rely on them.

References

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: "Automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Anastasopoulos, L. J. (2026). Artificial intelligence, bureaucratic discretion, and democratic administration. *Public Administration Review*. Advance online publication. https://doi.org/10.1111/puar.70201

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kotzian, P., Bauch, K. A., & Weißenberger, B. E. (2025). Accountability and managerial advice-taking: Comparing human and algorithmic advisers. *Management Decision*, *63*(13), 545–567. https://doi.org/10.1108/MD-07-2024-1618

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257

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