Philosophy & Technology (2026) 39:84
https://doi.org/10.1007/s13347-026-01090-9

RESEARCH ARTICLE

What is Wrong With Automation Bias?

Perica Jovchevski[1] · **Stefan Buijsman**[1] · **Mark Neerincx**[1]

Received: 16 December 2025 / Accepted: 27 March 2026

Abstract

This article examines the ethical and moral implications of automation bias in high-stakes decision-making contexts. Drawing on empirical studies, we distinguish between weak automation bias, where users follow system's automated cues (or its silence) without consulting readily accessible evidence that contradicts them, and strong automation bias, where users follow such cues (or their absence) even when they are aware of such evidence. While weak automation bias, in our view, resembles automation-based complacency and is plausibly associated with negligence on the part of the human operator, strong automation bias reveals an excessive and unwarranted transfer of trust from operators to automated systems which results in epistemic deference of the former to the prompts of the latter. We argue that what is ethically and morally troubling about this form of deference, is that it interferes with the exercise of the operators' autonomous agency as well as with their duty to exercise human judgment in high-stakes decision-making contexts. To mitigate these effects, we discuss two design-based tools introducing epistemic friction - Reflection Machines (RMs) and defeaters - which ultimately aim at cultivating critical trust in the interaction between human operators and decision-support systems.

Keywords Automation bias · Epistemic deference · Cognitive heuristic account · Overtrust account · Autonomy · Moral agency · Critical trust · Epistemic friction · Reflection machines · Defeaters

✉ Perica Jovchevski
p.j.jovchevski@tudelft.nl

Stefan Buijsman
s.n.r.buijsman@tudelft.nl

Mark Neerincx
m.a.neerincx@tudelft.nl

[1] Delft University of Technology, Jaffalaan 5, Delft, BX 2628, The Netherlands

Published online: 24 April 2026

1 Introduction

Human agents increasingly operate within socio-technical environments in tandem with automated or AI-enabled decision-support systems in both high and low risk decision-making processes. Medical professionals may consult diagnostic algorithms, pilots may supervise automated flight-deck technologies, a security analyst may rely on an AI assistant to summarize intelligence reports and users across different sectors rely on conversational agents and chatbots for information retrieval, decision guidance, or procedural support. In all these contexts, the outcomes depend not solely on either human expertise or machine computation capacity, but on the structure of their interaction. Understanding how AI systems shape human attention, judgment, and action is therefore critical, particularly in high-stake contexts where errors may carry significant consequences for individuals' safety, health, or even life.

It has been stressed in this respect that while these decision-support systems, through their alerting or recommending functions, are commonly designed to enhance human performance by guiding attention, reducing the cognitive workload or structuring better the decision alternatives, they also introduce new categories of risk that are related not merely to the possibilities of their technical malfunction but to the ways in which human operators perceive, interpret, and act upon their automated guidance. *Automation bias* exemplifies this class of risks, representing, more particularly, a tendency of human operators to over-rely on automated alerts or recommendations even in the presence of other indicators that in some way conflict or contradict them.

Empirical evidence of automation bias has so far been extensively documented across several domains. In their seminal aviation study, Mosier et al. (1992) demonstrated that when an automated checklist in a simulated engine-fire scenario erroneously advised pilots to shut down the engine that was not on fire, 75% of pilots complied with the faulty recommendation, disregarding contradictory cockpit indicators. In contrast, only 25% made the same error when relying on a traditional checklist. This discrepancy illustrates the extent to which automation can shape user behavior and override otherwise well-established procedural habits. In addition, their findings reinforce the conviction about the dual effect of automation: its capacity to enhance performance through precision and speed, but also its potential to amplify errors in cases of incorrect automated outputs. In the diagnostic use of clinical decision-support systems, it has been stressed that automation bias often leads to anchoring on system-generated differentials and treatment paths. Dratsch et al. (2023) have shown that erroneous recommendations from clinical decision-support systems significantly reduce diagnostic accuracy among radiologists of all experience levels when interpreting mammograms. Their findings further suggested that more inexperienced radiologists were more likely to follow the recommendation of the AI system, although this is by no means a conclusion to be generalized about the relation between experience and automation bias: for instance, in a later study Mosier et al. found that experienced pilots were more susceptible to omission errors than inexperienced ones (Mosier et al., 1998, p. 58). The costs of automation bias are perhaps among the most salient and tragic in the military domain. During the 2003 Iraq War, automation bias was implicated in incidents involving the U.S. Army's Patriot

missile system, which mistakenly engaged friendly aircraft—a British Tornado and an American F/A-18—resulting in the deaths of three aircrew members (Cummings, 2004, p. 6). Investigations later revealed that operators had acted in compliance with the system's automated threat classification despite observing that the system was unreliable, as some of the displays were incorrect.

Philosophers have examined how various forms of bias shape our reasoning and affect our moral appraisal of actions (Holroyd, 2012; Saul, 2013; Levy, 2016), producing a rich body of literature on implicit and cognitive biases as well as on algorithmic bias (Brownstein & Saul, 2016; Johnson, 2021; Peters et al., 2022). Yet automation bias has remained underexamined, despite its analogous structure and worrisome ethical implications. The present paper seeks to redress this imbalance by offering an integrated philosophical account of automation bias, one that draws upon empirical insights in articulating its conceptual specificities, its moral and ethical significance,[1] and the ways in which it may be mitigated or ameliorated through certain socio-technical policies and design tools.

Our analysis proceeds in three parts. In the first part, we outline two explanatory accounts of automation bias derived from empirical studies: we will call the first one the "cognitive-heuristic account," and the second the "overtrust account" of automation bias. The cognitive-heuristic account interprets automation bias as a byproduct of human cognitive economy, namely our tendency to use automated cues as heuristic substitutes for vigilant information seeking. The overtrust account, by contrast, explains automation bias as an outcome of disproportionate trust in automated authority, potentially driven by beliefs in the somewhat superior capacities of automated systems. Drawing upon the insights of the two accounts, we distinguish between weak and strong types of automation bias: weak automation bias arises when users follow automated cues, or the lack thereof, without consulting available evidence that contradicts them, while strong automation bias includes following automated cues despite attending to such evidence. We maintain that the two types of automation bias carry distinct ethical and moral implications, as they reflect different attitudes of the operator towards the accessible evidence and the automated outputs. Weak automation bias implies negligence on the part of the operator and, in high-stakes decision-making contexts, may constitute a moral failure to meet relevant standards of attentiveness and due diligence. Cases of strong automation bias, by contrast, involve epistemic deference to the system's cues, as operators treat those cues as having higher authority, thereby overriding or discounting other available indicators that point toward different courses of action.

Given that prior work has examined the moral significance of negligence in relation to weak automation bias and automation-induced complacency (Coco, 2023)—which share many features in common in our view—we focus in the second part of the paper on examining the normative significance of strong automation bias. We argue that, beyond the well-recognized long-term deskilling effects associated with all forms of automation bias, the deference characteristic of strong automation bias

[1] We distinguish morality and ethics along Scanlonian lines: morality is concerned with obligations we owe to other persons, whereas ethics concerns values related to how one ought to live one's life (Scanlon, 1998).

is ethically and morally troubling because it (1) interferes with the exercise of users' autonomous agency and (2) compromises their duty to exercise human judgment in high-stakes decision-making contexts. To address these concerns, in the final part of the paper we propose two design-oriented interventions which introduce epistemic friction in the interaction between the human operator and the decision-support system: Reflection Machines (RMs) and defeaters. Both of these tools aim at fostering reflective engagement and promoting critical trust in the decision-support system's recommendations and alerts, which in our view will contribute to reduction of the rates of both forms of automation bias across different domains.

2 Weak and Strong Automation Bias

In recent years, philosophers have developed two main accounts of the nature of bias: functional and norm-theoretic. On the functional view, biases operate as mechanisms for managing the uncertainty produced by underdetermination (Johnson, 2024) or as Johnson further specifies, they represent "non-evidential assumptions that systematically limit the inductive hypothesis space to a traceable size" (Johnson, 2024, 3/21). According to this approach biases are normatively neutral: they are not inherently objectionable and may in fact be indispensable for decision-making across a wide range of contexts. By contrast, the norm-theoretic account treats bias as "involving a systematic departure from a genuine norm or standard of correctness" (Kelly, 2022). The relevant norms here are norms of practical rationality, morality, and epistemic norms, and an important feature of this view is that the departures must be systematic rather than incidental. Both accounts are in principle compatible with pluralism concerning the things that can be biased—people, their beliefs, decisions, actions, the procedures through which they come to an outcome, and so forth.

The empirical studies on automation bias seem to instantiate, if not perfectly then to a considerable degree, this more general dichotomy about bias, developing two different accounts of automation bias which we will call the "cognitive-heuristic account" and the "overtrust account." One of the most elaborate and paradigmatic cognitive-heuristic conceptions of automation bias can be derived from the seminal study of pilot decision making in high-technology cockpits by Mosier et al. (1998), which defines automation bias as "omission and commission errors resulting from the use of automated cues as a heuristic replacement for vigilant information seeking and processing" (Mosier et al., 1998, p. 47). According to the authors, errors of commission occur when the user of a system actively follows an incorrect automated cue, even when other information is available that might contradict it. The example of pilots shutting down their engines in compliance with the system's false recommendation illustrates such a commission error. Errors of omission, on the other hand, occur when users fail to act appropriately because the system has not prompted them to do so, even though other available evidence indicates that they should have acted differently. For instance, a pilot commits an omission error when, relying on autopilot, he fails to correct an unnoticed altitude drift because the system gives no alert, despite clear countervailing cues such as a discrepancy between the altimeter and the flight-plan altitude. Underlying both types of error is what Mosier et al. call a replace-

ment of vigilance: users substitute active information-seeking and cross-checking behaviors with acceptance of an automated cue, in the case of commission errors, or the absence thereof, in the case of omission errors.

Several further aspects of this conception, elaborated in subsequent work by Skitka et al. (2000), are crucial for understanding the nature of automation bias. First, they characterize it as a decision bias arising from cognitive appraisal and processing limitations, rather than from any particular belief in the relative authority of automated systems (Skitka et al. 2000, p. 704). On this view, automation bias stems from using automation as a heuristic substitute for more vigilant and comprehensive information processing. The support for this claim is grounded in the cognitive miser hypothesis, according to which humans are disposed to minimize cognitive effort when making decisions (Kahneman et al., 1982; Fiske & Taylor, 1994). A second important feature implied by this conception is that automation bias presupposes the presence of evidence contradicting the automated cue. It is only when human operators have access to such conflicting information that one can meaningfully identify their response as biased, typically by observing whether there was or was not verificatory behavior (Skitka et al. 2000, pp. 705–707). The presence of contradictory or conflicting evidence is thus a conceptual feature of automation bias—common to both the cognitive heuristic account and the overtrust account discussed below—which on one side is constitutive of the concept of automation bias and on the other helps us with the practical epistemic task of identifying something as automation bias.[2] Note that, based on this feature, the same commission or omission error in the use of a decision-support system might be classified differently. For example, a pilot who follows an automated navigation instruction despite visible discrepancies on the flight instruments might exemplify automation bias, whereas one who acts in the same way without any contradictory evidence available does not.

It is important to emphasize that the degree to which automation bias occurs is mediated by a variety of factors. According to Parasuraman & Manzey (2010), these are related both to system properties and to the task context. Bias tends to increase when automated aids issue specific action recommendations rather than merely supporting analysis, and when users lack trial-by-trial confidence feedback. Task-related factors, on the other hand, include the distribution of task accountability, perceived criticality of events, workload, time constraints, and individual differences. Clearly, some of these task-related factors may be such as to place a commission or omission error, even in the presence of contradictory evidence, outside the circumstances of automation bias. For instance, cases in which such errors occur due to an operator being tasked with an unreasonable workload can be considered to fall outside the concept of automation bias. [3] Finally, we should also stress that while traditional

[2] In some accounts such as Cummings (2004) this feature is even more pronounced. She defines automation bias as occurring "when a human decision maker *disregards or does not search for contradictory information* in light of a computer-generated solution which is accepted as correct." (Cummings, 2004, p. 2; Emphasis added).

[3] Recently some have observed for instance that the use of the Lavender AI-system by the Israel Defense Forces (IDF) during the Gaza War instantiates automation bias (Downey, 2025). This decision support system processed large volumes of data at high speed to produce lists of potential targets. Evidence suggests that human analysts frequently approved the AI-generated targets with minimal scrutiny, sometimes

human factors accounts have framed automation bias primarily as a multitasking phenomenon driven by divided attention, the evidence synthesized by Lyell and Coiera (2017) demonstrates that automation bias also emerges in single-task environments, particularly in medical and diagnostic contexts. Their review shows that, for instance, tasks such as radiological cancer detection, computerized ECG interpretation, and clinical decision support are susceptible to automation bias even without competing secondary tasks.

Aside from the cognitive-heuristic view, other empirical studies suggest a somewhat different understanding of automation bias as primarily resulting from overtrust in automated decision aids. [4] In their integrated account, Parasuraman & Manzey (2010, p. 406) consider automation bias—along with complacency—to be "phenomena that describe conscious or unconscious responses of the human operator induced by overtrust in the proper function of an automated system." [5] While there is considerable variety in understanding trust in automation as well as overtrust, the latter is generally understood as a psychological, attitudinal and behavioral state of miscalibrated trust—in the case of automation bias, users placing more trust in an automated system than is warranted by its actual capabilities and reliability (Ullrich et al., 2021, 2/15).

Empirical research on overtrust and overreliance shows that there are at least two groups of prominent factors that contribute to overtrust in an automated system, some

dedicating only a few seconds to each (Abraham, 2024). Given the insights into automation bias above and to the extent it is unreasonable to suppose that the operators of these system had sufficient time for verification or for acting on a verified information to the contrary, the killings due to these errors cannot be claimed to be due to automaton bias. They are rather examples of wrongful use or malicious misuse of decision support system.

[4] Overtrust, and the resulting overreliance, is widely recognized as a primary cause of automation bias across discipline such as ergonomics, legal studies, medicine etc. (Romeo & Conti, 2025; Coco, 2023; Godard et al.2012; Schemmer et al. 2022). Skitka et al. (2000) also found that there was some correlation between believing in the decision-making superiority of computers and the occurrence of commission errors, but they argued that this relationship was relatively weak (Skitka et al. 2000, p. 714). In earlier work by Skitka et al. such beliefs strongly predicted compliance with computerized recommendations that directly contradicted available information. However, in that case, all relevant data were easily accessible, requiring little vigilance. In their view, when information becomes harder to obtain, however, factors such as attentional vigilance and cognitive effort, or their absence, play a far greater role in determining reliance on automation than do beliefs in its epistemic authority.

[5] It is interesting that with only one exception, none of these studies of automation bias as overtrust mentions the prevalence of the opposite tendency of under-trust in the automated cues, as a kind of "negative automation bias". Parasuraman & Riley (1997) do note this phenomenon but consider it a "disuse" of automated cues. In their view disuse occurs when operators fail to utilize automation that is superior to manual methods, due to lack of trust in the system, which is, however, more reliable then they believe. Recent empirical research has increasingly examined this opposing tendency under the label of algorithm aversion. This literature documents cases in which individuals systematically discount, avoid, or override algorithmic advice after observing or experiencing system errors, even when the algorithm demonstrably outperforms human judgment (Dietvorst et al., 2015). Unlike automation bias, which reflects overtrust, algorithm aversion captures a form of excessive skepticism toward automated cues. The coexistence of these tendencies highlights that trust miscalibration in human–automation interaction can manifest in both directions: as unwarranted reliance or as unwarranted rejection to rely. Acknowledging the tendency of algorithm aversion is therefore important for avoiding overgeneralized claims about persistent overreliance on automation, while remaining consistent with the substantial evidence that automation bias constitutes a robust and recurrent risk across many domains.

of which are related to the user's personal experience with the system, while others are related to the systems' reputation as capable and reliable. It has been well documented that human operators—either based on the positive reputation of these systems or based on a positive experience in interacting with automated systems—tend to perceive and treat them as highly capable agents, sometimes even as possessing decision-making abilities superior to their own or those of other humans (Lee & See, 2004). Empirical support for these claims can be found in the study by Dzindolet et al. (2002), in which participants were asked to predict the performance of an automated system relative to a human assistant in a decision-making task. The results revealed that most participants initially expected the automated aid to outperform the human aid by a substantial margin, illustrating a strong bias in favor of perceived automated competence. Of course, trust calibration is dynamic, and what constitutes initial trust may change during the use of the system depending on the system's reliability and the user's understanding of how the system operates (Dzindolet et al., 2003, pp. 703–704). If operators observe that the system makes mistakes, their trust in it may decline. However, as Dzindolet et al. (2002) found, offering an explanation for why an automated aid might err increased users' trust in the system even though participants in their study were informed that the automated system had made twice as many errors as they had. This study demonstrates not only that there is initial level of more trust in automation than in humans for certain tasks that both can perform but also that such trust can easily be miscalibrated into overtrust. We should also add to this that, overtrust in automation may be significantly shaped not only by the reputation and perceived capability and reliability of the system but also by a combination of other system-related factors such as the design of the cues, the opacity of the automated processes, as well as factors belonging to the socio-technical system more generally—for instance, institutional norms that discourage challenges to automation.

Note that, unlike the cognitive heuristic account, the overtrust account need not necessarily imply that biased individuals fail to attend to easily available evidence indicating a different course of action than the automated cue. The overtrust hypothesis is compatible with an operator acting on the cues of the system, or the lack thereof, even if he has attended to evidence that contradicts the cue. This is so because it is possible that he may still treat the automated system as more trustworthy than other indicators or one's own judgement for any of the reasons mentioned above which make such trust miscalibrated into overtrust. This difference between the two accounts is important because the way users relate to accessible evidence that contradicts an automated cue is normatively significant: as we discuss below, it manifests different types of not only epistemic but also ethical and moral failures. At this point a legitimate question arises: what counts as accessible evidence here? Is it perceptual awareness, procedural availability, or inferential accessibility to such evidence? In many high-stakes contexts, contradictory evidence may be technically available but not practically accessible. So a plausible interpretation of the "access to conflicting or contradictory evidence" condition would in our view be that there is inferential access, namely information that could be recognized as contradictory with optimal cognitive effort given role-appropriate expertise. It is this kind of access that becomes ethically and morally salient. By contrast, when contradictory information is only procedurally or technically available but practically inaccessible because of

unreasonably high costs for the operator – which might be due to design, workload, etc. - this condition is not met.

Based on how the operator relates to this accessible evidence, we would like to propose a distinction between two types of automation bias: weak and strong. Weak automation bias results from a failure to attempt verification of an automated cue,[6] whereas strong automation bias persists even after attending to contradictory evidence. For illustration, consider the following two examples of weak and strong automation bias, respectively. Lyell et al. (2017) investigated the phenomenon of automation bias in the context of electronic prescribing, where 120 final-year medical students used a simulated e-prescribing system with clinical decision support (CDS) under varying conditions of task complexity and interruption. They found that, compared to a no-CDS control condition, scenarios with incorrect CDS advice led to a 33.3% increase in omission errors and a 65.8% increase in commission errors in low-complexity scenarios, where one of the reasons for the commission errors is cited the reduced sampling of information that could verify the decision support. This failure to verify is taken as an indicator of what we defined as weak automation bias.

By contrast, the study by Rosbach et al. of AI-assisted diagnostic decision-making in computational pathology closely matches what we term strong automation bias. In their experiment, 28 trained pathology experts estimated tumor-cell percentages (TCP) both with and without AI assistance, under conditions with and without time pressure. Of 560 AI-assisted TCP estimates, practitioners adopted AI recommendations contradicting their independent assessments in 67 cases. Of these, 29 were positive consultations, where reliance on the system improved accuracy, while 38 were negative consultations, in which initially correct evaluations were overturned by erroneous AI advice. It is these latter consultations that instantiate a miscalibrated trust, or over-trust in the AI system that amount to 7% ration of automation bias (Rosbach et al., 2025). Importantly, time pressure did not increase the occurrence of automation bias, which remained stable across conditions, but it amplified its severity: when experts followed incorrect AI recommendations under time constraints, performance degradation was substantially larger and alignment with AI advice increased. This pattern is consistent with our account of strong automation bias in which agents defer to automation despite having formed correct independent judgments, and where cognitive load factors modulate the magnitude rather than the frequency of biased reliance (Rosbach et al., 2025).

While weak automation bias is widely regarded as psychologically plausible, the very notion of strong automation bias might invite skepticism.[7] First, one may question whether studies such as Rosbach et al. (2025)— which do not explicitly control for operators' confidence—truly capture strong automation bias, rather than a form of rational deference to the AI system driven by insufficient confidence in one's own capacities, estimations, evaluations, beliefs. If the latter is the case then these instances may be more appropriately interpreted as cases in which operators lack

[6] This characterization of weak automation bias, however, reveals overlaps between weak automation bias and automation-based complacency—understood as reduced detection of system malfunctions under automation compared with manual control (Parasuraman & Manzey, 2010, p. 390).

[7] We are thankful to an anonymous reviewer for raising this objection.

adequate confidence in their independent judgments and often based on prior positive experience with the AI system, assign greater trust to its outputs than to their own assessments. Since relying on the AI system in these cases seem rational - the objection goes - they would clearly not qualify as cases of bias at all. What would be left then for the notion of strong automation bias would only be cases in which one is sufficiently confident or rather certain that the system's cue is incorrect, nevertheless follows it. However, in absence of control for confidence in experiments such as the one above, a skeptic might claim that it is highly implausible that such situations exist. And if no clear cases can be identified in which an individual is sufficiently confident or certain that the automated recommendation is incorrect yet still relies on it, then the category of strong automation bias may appear not only empirically rare but even conceptually unstable.

The problem with this objection is that it unjustifiably narrows the scope of strong automation bias, by implausibly narrowing down the scope of cases that count as overtrust and overreliance. Two types of reasons make this narrowing down implausible: conceptual, related to the concept of overtrust and substantive, related to cases we actually judge to be biased. As we mentioned above, conceptually, overtrust in a system refers to a mode of miscalibrated trust. The empirical literature on trust in automated systems considers trust to be appropriately or well-calibrated when reliance to a system tracks the system's actual capacities, competence, and situational limits and it is miscalibrated when reliance exceeds what those properties warrant (Muir, 1987; Lee & See, 2004; Hoff & Bashir, 2015; Kraus et al., 2019). So calibrated trust is an optimal state reached in the interaction with a system where the trust invested in the system exactly matches the objective capabilities and performance of the system. Further, we find whether trust in a system is well-calibrated or not, when by isolating other variables that might contribute to reliance – such as time pressure, for instance - we evaluate the outcomes of the interactions between the user and the system: as in the Rosbach experiment above, we consider whether they resulted in correct or incorrect decisions. The level of confidence a user has in one's own assessment of possible courses of action does not play a role in determining trust calibration at all. Of course, the level of self-confidence is a causal factor which influences the *propensity* of the operator to trust, but it is not constitutive of trust calibration.[8] Accordingly, it cannot serve as a necessary condition for diagnosing overtrust in an automated system in the first place, nor consequently, for diagnosing bias.

Note further that there are also substantive reasons for rejecting this objection. Even if one grants that trust calibration should in some sense be confidence-sensitive, it does not follow that any deference under conditions of insufficient trust of the operator coupled with investment of more trust in the system than in one's own capacities would be rational or non-biased. Imagine a loan officer evaluating a small-business applicant whose financial record displays atypical but legitimate income variability. Suppose that the officer judges the applicant as "likely creditworthy" and assigns

[8] As we already stated above many studies have investigated in this respect the presence of automation bias in inexperienced, moderately experienced and highly experienced practitioners and found that all groups are prone to it, with different studies having different results about which of these groups are more susceptible to commit commission and omission error (Dratsch et al., 2023, Mosier et al., 1998).

confidence level as 2/4 (“not confident”) to his estimation. He consults an AI scoring system which assigns a high-risk label. It is known to the officer that the AI system bases its recommendation on patterns correlated with default in a dataset which however underrepresent gig-economy profiles. Nevertheless, the officer defers to the AI system and records the reasons as being related to his general positive experience with the predictive accuracy of the AI system. If his decision proves to be incorrect it seems that we are justified in claiming that his decision was biased, irrespective of his trusting the system more, because given his knowledge of the limitations of the system, he had good reasons not to invest that level of trust in the system, in this particular case. If this is so, then we can conclude that self-confidence neither functions as a necessary condition for determining overtrust nor, if it was such a condition, it would have necessarily make deference to the system’s cues rational or non-biased. This suffices to reject the objection above and secure the existence of the class of cases that manifest strong automation bias irrespective of the confidence level of the user. Recent empirical research on trust in AI systems offers direct support for such cases. For instance, Klingbeil et al. (2024) implemented a modified six-round trust game with fixed roles and stranger-matched dyads, in which Player A, after forming an independent assessment based on Player B’s record of prior IN/OUT choices, received in the fourth round an advice framed as AI- or expert-generated before deciding whether to cooperate (IN) or not (OUT) with the player. Their results indicate that merely labelling advice as AI-generated significantly increased reliance. For example, among participants who independently judged Player B “very likely OUT,” cooperation rates rose from 6% in the control group to 44% under AI advice encouraging cooperation, whereas expert advice produced no significant effect. Their experiment gives strong support to the existence of cases of strong automation bias as it explicitly tests the thesis that users follow the AI advice even when it contradicts available contextual information as well as their own assessment.

Why is it importnant to make this distinction? As mentioned above, the reason to distinguish these two types of bias based on the engagement of the agent with accessible contradictory evidence is that they seem to have different normative significance. Weak automation bias is considered “weak” because it arises from minimal cognitive engagement: users follow automated cues, or the lack thereof, largely through neglect rather than through any active assessment of the situation. This type of negligent actions exhibiting automation bias in high-stakes contexts is morally troublesome as it involves a culpable failure to exercise the level of care that a situation reasonably demands, thereby exposing others to avoidable risk or harm. Its wrongness does not stem from an intention to cause harm by the agent but from the fact that the agent ought to have known better – given the access to such knowledge and one’s role-based duties - and acted with greater caution. By contrast, strong automation bias is “strong” because it persists even under conditions of active engagement with evidence to the contrary. Here, users do attend to countervailing evidence and engage in some deliberation—which can range from simple registration of the evidence to weighing of evidence and forming their own assessment on the case — yet still accord the automated outputs higher trust or epistemic weight. In respect of the way operators may be attending to the evidence, Parasuraman notes that there is limited empirical evidence about whether commission errors genuinely reflect a bias

in how information from different sources (automation versus one's own sampling) is weighted, or whether they indicate further neglect of the evidence attended (Parasuraman & Manzey, 2010, p. 398). Both seem to be plausible options given the insights of the empirical studies discussed above. Unlike weak automation bias, which implies culpable negligence, we will argue in what follows—on the ground of the overtrust hypothesis—that strong automation bias entails epistemic deference,[9] by which we mean treating another agent's judgment on a matter as more authoritative or credible than one's own simply because it comes from that agent. [10] This deference, we will argue, constitutes not only an epistemic but also an ethical and a moral failure.

3 The Ethical and Moral Concerns Arising From Strong Automation Bias

Scholars have long warned that automation can weaken the quality of human decision-making, impair monitoring performance, and reduce the likelihood that operators will intervene when automated outputs are flawed. Perhaps among the most concerning effects of automation, extensively documented across aviation, medicine, and other high-stakes domains, are deskilling—the erosion of previously acquired competencies through disuse—and its counterpart, upskilling, the suppression of opportunities to develop or maintain advanced expertise due to heavy dependence on decision-support systems (Natali et al., 2025, p. 356). Automation bias adds to these an additional layer of ethical and moral concern stemming from automation. As argued earlier, weak automation bias can amount to negligence, which can be morally worrisome, as in high-stakes environments failures to attend to countervailing evidence can have life-or-death consequences and are prima facie culpable. In this section, we examine, however, the rather unexplored aspect of the distinctive ethical and moral implications of strong automation bias. In what follows, we will argue that the ethically troubling aspect of strong automation bias lies in how the deference it implies can compromise the autonomous agency of operators (2.1), while its primary moral concerns are associated with its interference with the duty of operators to exercise human judgment in high-stakes decision-making contexts (2.2). This latter argument, to some extent, also applies to weak automation bias, although our primary motivation is to explain wrongs specific to strong automation bias.

To be clear at the start, we do not argue that epistemic deference to automated cues is inherently wrong. In some cases, such deference may be rational, particularly when systems reliably and significantly outperform human cognition in tasks that humans cannot verify or that would require an unreasonably long time to assess. In these circumstances, human intervention might even be undesirable. The problem arises,

[9] For empirical studies treating automation bias in general as implying deference see Alon-Barkat & Busuioc (2023) and Zerilli et al. (2024).

[10] The concept of deference has different elements depending on the context of use, some of which we incorporate in our definition. See more on the different elements of deference in the debates over expert deference, deference to authority, deference to marginalized groups in Palmira (2020), Davia (2015), Bokros (2021), Tilton & Toole (2010).

however, when, as in strong automation bias, operators have attended to or weighed contradictory evidence based on which they have assessed or can assess the automated cue as incorrect, yet nonetheless follow the system's recommendation or alert.

It is important to also recognize that not all forms of deference to decision-support systems arise from strong automation bias. Institutional structures can produce morally and ethically problematic forms of deference independently of an individual's personal experience or trust in automated systems. For example, Gravett (2023, 291) cites a survey of Canadian judges and lawyers showing that, although many did not regard risk-assessment tools like COMPAS as particularly reliable for predicting future behavior, they nonetheless preferred using them because doing so reduced the personal and professional risks associated with their decisions. Similarly, in healthcare, Pozzi (2023, 5–12) illustrates how physicians who wish to prescribe outside the NarxCare system's recommendations are exposed to pressures that may flag them as "overprescribers," risking reputational harm and even professional sanctions, including potential loss of license. While these forms of deference are morally and ethically concerning, they do not originate from automation bias itself; rather, they reflect institutional incentives that, over time, can foster uncritical thinking and undermine evaluative engagement.

3.1 Compromising Autonomous Agency

Strong automation bias, as defined above, can be said to interfere with the exercise of autonomous agency by operators of decision-support systems in at least two distinct ways, depending on the conception of autonomy one adopts—namely, whether one presupposes procedural or substantive conception of autonomy. There seem to be two proceduralist arguments which support the compromised autonomy claim: the first one relies on a claim that deference to the system might imply internally manifested inconsistency of beliefs, while the second one relies on a claim that such deference may undermine individuals' capacities to revise their own beliefs. Both proceduralist arguments are comparatively weaker than the substantivist one in that they either hold only for a subgroup of cases of strong automation bias or for the whole class of such cases but only under certain conditions. The substantivist argument, on the other hand, builds more directly on the previously discussed considerations regarding overtrust and epistemic deference to automated cues. It relies on the claim that the socio-technical cultivation of such overtrust and its consequences attenuate the epistemic and normative resources necessary for individuals to regard themselves as legitimate decision-making authorities in their interactions with decision-support systems, thus interfering with their exercise of autonomous agency. Before proceeding with the arguments, let us specify the mode of autonomy under consideration. Following Feinberg, we may distinguish, at least, four modes or senses of autonomy: as a capacity for self-governance; as a condition of agents who are self-governed (or of their actions); as an ideal of agency; and as a form of personal sovereignty protected by rights (Feinberg, 1986, 27–51). The arguments we present are ultimately concerned with autonomy understood as a condition of agents' actions, although in both cases we will also make reference to the capacities for such actions.

3.1.1 The Proceduralist Argument

Agents act autonomously when they guide their actions in light of beliefs, rules, principles, or values that are, in some sense, "their own" (McKenna, 2005, p. 207). Although they appear in vast diversity, we can say that in general, according to proceduralist accounts of autonomy, what makes an agent's action autonomous is determined by how the desires or beliefs from which the action arises have been formed and how they fit within the mental ecology of the agent. Proceduralists typically distinguish three types of conditions for autonomous agency. First, a capacity condition, which refers to the possession and exercise of reflective capacities, including minimal rationality, self-awareness, and other related abilities (Christman, 1991, 13–16). Second, an endorsement condition, which can take various forms, such as the identification of higher- and lower-order desires or the absence of alienation from one's desires (Frankfurt, 1988, 19–22; Christman, 1991, p. 11). Third, a procedural independence condition, which protects the endorsement process from influences that subvert reflection, such as subliminal cues, manipulation, or coercion (Dworkin, 1988, 18). Our argument below focuses on the rationality and the capacity requirements, so it is worth clarifying only the content of the capacity condition. While there are different views about how to conceive rationality in autonomous action, virtually all proceduralists agree that agents should have an internally consistent set of beliefs. However, strict consistency is clearly too demanding a condition, as probably all of us hold inconsistent beliefs at some level. So, this condition is mainly understood as a minimal rationality requirement demanding that agents' actions are not guided by what Christman calls manifestly inconsistent desires or beliefs, namely those desires or beliefs "which the agent could bring easily to consciousness and recognize as incompatible" (Christman, 1991, p. 15).

With these theoretical suppositions about autonomy in place, let us consider a somewhat modified version of the example of strong automation bias above. Imagine that an experienced radiologist uses a clinical decision-support system that automatically flags lung nodules as benign or suspicious. During a review, the system classifies a particular nodule as benign (let us stipulate this as action x: "discharge the patient without follow-up"). At the same time, the radiologist observes two salient indicators that ordinarily warrant further investigation (let this set of reasons imply that one should do action non-x): the nodule has irregular spiculated edges, and the patient has a significant smoking history. Suppose further that the radiologist knows (with sufficient confidence or certainty) that both indicators are well-established red flags in the diagnostic literature. If the radiologist judges, based on these independent clinical markers, that the nodule is likely not benign (i.e., which implies he should do non-x), yet still defers to the system by doing x—because, he holds a belief that the system is more authoritative on the issue than himself—then doing x implies a violation of the internal consistency requirement for his action. His performing action x does not align with his belief implying non-x. Under procedural accounts of autonomy, this mismatch indicates a breakdown in self-governance, resulting in non-autonomous action generated by strong automation bias.

Note however that this proceduralist argument does not hold for all cases of strong automation bias. In the example above, the argument relies on a premise that the

user of the decision-support system is sufficiently confident of certain that the two other indicators are red flags and yet acts against this knowledge. It is true that in some empirical studies on automation bias their design was such that this condition was met. For example, in Skitka et al. (2000)—which, however, examines what we defined above as weak automation bias—participants were explicitly informed that the automated system was highly, but not perfectly, reliable, while the other indicators were fully reliable (Skitka et al. 2000, 706). However, this condition cannot be generalized across the whole class of instances of strong automation bias. It is possible that there are cases in which the operator is simply undecided or insufficiently confident and having higher level of confidence or trust in the system, defers to the system's recommendation. Here, even if it is the case that the operator judges that non-x, this may not in itself be a sufficient reason for him to do non-x, because he trusts the system's cue more than he trusts his own decision-making powers. If this latter belief is something he autonomously possesses then it seems that his action follows his rational deliberation. While cases like this might imply deference due to overtrust they will not imply violation of the consistency requirement for autonomy. So in this sense the argument above is limited in scope.

The proceduralist has, however, several other resources among the capacity conditions for autonomous action that might accommodate these cases. Dworkin has persuasively argued that autonomy implies an ability to scrutinize critically one's motivations, desires, preferences, and also to change them if one wants to (Dworkin, 1988, 16–17). Something similar holds for the autonomy with respect to our beliefs too, with the difference that changing our beliefs is not a matter of us wanting to believe something or not; it is a matter of correspondence to facts, which is independent of our desires. In this respect, in order for our beliefs to be autonomous, it would be required not only that we retain a capacity to form them, but also a capacity to revise them and change them where appropriate. If this is so, one may ask—in respect of the cases where there is insufficient confidence in one's capacities but more trust in the capacities of the system—to what extent the beliefs on which this trust is based are possessed autonomously in the first place? The beliefs that give rise to the perceived reliability, authority, and trustworthiness of the system are second-order beliefs that hinge on first-order beliefs about the system's capacity to offer correct cues: the more individuals observe that the system's cues are correct, the more reliable they consider the system to be, the more authoritative on particular epistemic matters, and the more trustworthy.[11] Similarly, in light of first-order reasons that the system's automated cues might not be correct—as is the case when there is evidence that contradicts them—rational individuals are expected to at least weaken their trust in the system's capacities. [12] So, a first-order belief about the incorrectness of the system's cues should be reflected accordingly in the second-order belief about the system's reliability, authority, and trustworthiness. When this transitivity does not hold, then we seem to have good reasons to suspect that the trust in the system functions as a non-autonomous element in the mental economy of the operators. If this

[11] To be clear, we do not claim that reliability is all that makes up the authority or trustworthiness of a system, but we do believe it is the central part of it.

[12] Empirical support for this dynamic of system trust recalibration can be found in Dzindolet at al. (2002).

is so, then a proceduralist argument might be crafted that demonstrate how cases of strong automation bias undermine the autonomous agency even in cases where there is insufficient confidence of the agent in one's own assessment.

We need to point out, however, to the limits of this argument. One may question whether an agent's failure to reduce trust when confronted with counterevidence necessarily indicates that such trust is non-autonomous, or at odds with procedural self-governance.[13] From a procedural perspective autonomy does not require flawless or immediate conformity to epistemic norms. An agent may reasonably regard a specific system output as mistaken yet still maintain the view that the system is, overall, more reliable than his own judgment, particularly in probabilistic environments where occasional errors are expected.

While this objection rightly notes that an agent's global assessment of a system's trustworthiness need not be weakened by a single erroneous output — and that autonomy may therefore remain intact — it overlooks that such insensitivity is normatively constrained within a rational range. Autonomy does not require agents to revise their beliefs mechanically in response to every isolated error. Yet it equally does not license the persistent or systematic disregard of relevant countervailing evidence. The central issue, then, is not whether trust calibration shifts after one mistake, but whether the agent remains appropriately responsive to the evidential significance of errors once they become recurrent. Although this requirement may appear to imply a thicker epistemic condition than procedural accounts of autonomy would typically endorse, it need not be so, as it can be grounded in the capacity condition itself, specifically in the capacity for belief revision. Possessing such a capacity presupposes a degree of sensitivity to evidence and we partly infer its presence by examining how agents treat evidence bearing on their beliefs. When agents repeatedly abandon their own initially correct judgments in favor of incorrect system outputs, the resulting pattern of error cannot always be explained through rational deference under conditions of low confidence and the investment of superior capacities to the system. Beyond a certain threshold, continued deference without corresponding adjustment of trust becomes difficult to justify. And where this adjustment fails to occur, we acquire reason to question whether the agent's capacity for belief revision is being adequately exercised.

So unlike in the cases in which there is sufficient confidence or certainty on part of the agent in his own assessment, where we can infer whether an action is autonomous or not, in cases of insufficient confidence that might indeed not be the case. Nevertheless, there are identifiable thresholds beyond which continued deference despite accumulating counterevidence may signal deficiencies in one's capacity to revise one's beliefs, and consequently to one's capacity to act autonomously.

3.1.2 The Substantivist Argument

Substantive conceptions of autonomy are committed to the claim that the possession of some kinds of beliefs (Stoljar, 2000) or the lack of some self-regarding attitudes (Mackenzie, 2008) might also make an action non-autonomous, aside from the action

[13] We are thankful to an anonymous reviewer for raising this objection.

meeting all procedural conditions. It is interesting to note that at the place where the second proceduralist argument ended, a substantivist argument for the claim that strong automation bias interferes with the exercise of autonomy of the operators can begin to develop. According to this argument, the deference of the agent to the system's cues need not reveal inconsistency in the biased operator's belief system which in turn undermines the autonomy of their actions. Rather, when one defers to the system's cue aside from attending to evidence to the contrary, that signals a lack of attitude towards oneself as a legitimate authority to make decisions concerning the tasks that are done in interaction with a decision-support system. The lack of such self-regarding attitude makes one's action non-autonomous.

The substantivist argument starts from the premise that automation bias is not merely endogenous to the psychology of agents but is structurally and technically induced through the distribution of epistemic authority within the socio-technical system. Although different socio-technical systems may embed different patterns of human–system interaction or collaboration with different authority distributions, it is common to all that they presuppose somewhat superior capacities of automated systems in dealing with particular tasks. Several features of the socio-technical system can be invoked in support of this claim. The first one is the social perception and acceptance of the purported objectivity and neutrality of system outputs. These outputs are typically generated by formalized algorithms, data analysis, or "evidence-based" procedures, and are often regarded as systematic and consistent forms of decision-making. This social perception contributes to treating system outputs as having a higher level of reliability and objectivity, granting the system a special epistemic standing vis-à-vis the human operator within the socio-technical context. [14] A second feature concerns the growing prevalence of organizational requirements that structure how system outputs are to be used. These requirements operate through at least two distinct pathways. First, certain institutional arrangements disincentivize contestation of system outputs. When operators are required to provide extensive justification for deviating from a system's recommendation, disagreement becomes procedurally costly. Similar asymmetries may arise through audit practices, performance evaluations, and so on. Although operators may retain nominal decision authority, the justificatory architecture frames the system's output as the default standard against which human judgment must be defended. The predictable effect is not the elimination of contestation, but its gradual repositioning as exceptional, and risk-laden. Over time, agents may become less willing to challenge system outputs, not because they lack reasons, but because institutional norms systematically increase the friction associated with exercising independent judgment. This dynamic ultimately may weaken the practical conditions under which operators can enact and sustain their status as credible epistemic authorities within the system. Second, as noted above, persistent requirements to consult or rely upon automated systems can generate a deskilling dynamic. Where system guidance is continuously foregrounded, opportu-

[14] In support of this stands the study by Dijkstra et al. (1998), which found - based on an experiment with 84 college students - that, when presented with identical reasoning, the students treated the advice attributed to an expert system as more objective and rational than the same advice attributed to a human advisor. These perceptions suggest further that users' beliefs about computer objectivity can significantly shape how they engage with and rely on expert systems.

nities to practice diagnostic discrimination, evidential weighing, and error detection may be reduced. The concern here is not merely diminished technical proficiency but the attenuation of capacities constitutive of professional epistemic agency. As these capacities are exercised less frequently, agents may experience a decline in self-trust, namely in confidence in their ability to form, assess, and defend judgments without system mediation. Taken together, these mechanisms exert a mutually reinforcing effect. Disincentives to contestation constrain when and how independent judgment is expressed, while deskilling alters the underlying competence and self-confidence required to sustain such judgment.

Lastly, in many operational domains such as diagnostics, navigation, and risk prediction, automated systems demonstrably outperform average human performance on routine, high-frequency, or data-intensive tasks. This empirical success produces a feedback dynamic: observed reliability becomes evidence supporting further reliance, which in turn amplifies the system's practical and epistemic authority. Over time, this habitual reliance may tilt the balance of epistemic authority in the system's favor, even in contexts where human judgment might be necessary and decisive.

Based on these insights the substantivist argument concludes that the deference of the operator in cases of strong automation bias may indicate a socio-technical system that *attenuates* the epistemic and normative resources necessary for operators to regard themselves as a legitimate authority in decision-making. Thus, unlike the proceduralist argument, this argument treats the exercise of the autonomy of human operators as institutionally scaffolded and a relational phenomenon. Strong automation bias is not merely an individual failure but rather represents a distortion in the distribution of epistemic authority within socio-technical systems, which compromises the conditions under which individuals can be seen as legitimate sources of autonomous judgment on matters on which they collaborate with a decision-support system.

3.2 Interference With the Duty to Exercise Moral Agency in High-Stakes Contexts

Although we increasingly rely on automated cues and decision-support systems in high-stakes decision-making environments, it is widely accepted that decisions involving matters of life and death or which might have grave consequences for the well-being of affected persons should not be fully entrusted to such systems. Perhaps this worry about entrusting systems to make such decisions has been most elaborately discussed in the debates over the use of autonomous weapon systems in the context of armed conflict. The core of what has come to be known as the dignity-based objection (Heyns, 2017; Sparrow, 2016) to the use of such systems rests on the conviction that human life possesses inherent worth that demands recognition and concern, even in the adversarial context of war. While combatants and non-combatants may forfeit or endanger certain rights, they nonetheless retain their status as moral beings whose humanity cannot be alienated, objectified, or instrumentalized for strategic ends. This intrinsic value imposes constraints on how force may be used: even lethal use of force must remain bounded by moral considerations that acknowledge the moral standing of those affected. They must, in some sense, be treated as ends in themselves and not merely as means to the realization of operational objectives or military advantage.

Autonomous weapon systems, however, are incapable of appreciating this value in any meaningful way. They cannot recognize the moral significance of their actions or show genuine concern for the people who may be killed as an effect of those actions. Hence, "delegating" the task of killing to these systems in armed conflicts represents a lack of concern for the humanity of the people involved in the armed conflict and as such amounts to a disregard of our commitment as humans to the principle of respect for human dignity. From this perspective, the conclusion of the debate concerning autonomous weapons is that human control over autonomous systems must ultimately rest upon the exercise of human moral judgment. Such control should always be preserved—in some meaningful form—within the so-called "kill chain." This preservation is essential, first, to ensure that decisions involving lethal force are subject to moral appraisal, and second, to affirm the dignity of all human beings who may be targeted or affected by those decisions. The inclusion of a human operator "in" or "on" the loop is thus intended not merely to meet procedural or legal criteria, but to guarantee the presence of genuine moral agency and the exercise of human moral judgment in the decision process.

The question, then, is whether strong automation bias compromises this very exercise of independent human judgement, even when a human remains formally "in" or "on" the loop. We argue that it does. It seems, in fact, that the outcome of an error arising from deference, as implied by strong automation bias, is morally more troubling than the same outcome resulting from a fully autonomous system that was explicitly delegated the authority to make life-and-death decisions. Although the human's presence within the loop is necessary to ensure that the dignity of those affected by the system's operation is respected, it is nevertheless not sufficient. Human agents occupying this role are required not merely to act as "rubber stampers" of the system's recommendations, but to actively exercise a role as moral agents when acting upon them.

This role implies that they are morally accountable or answerable for their actions, namely they are bound by a duty to justify them in light of first-order reasons that apply to particular cases and bear on the moral permissibility of their actions. When they do so they express appropriate moral concern for those affected by their decisions. The deference implied by strong automation bias makes it difficult for agents to meet these justificatory demands. When agents defer to automated recommendations despite evidence that contradicts them, they replace first-order reasons bearing on the moral permissibility of their actions with second-order reasons grounded in their belief in the system's reliability. Consider the justification an agent might offer in such a case: "There was evidence that the system's output was incorrect (or I assessed it as incorrect), but I still trusted the system because it has generally proven reliable in the past." In high-stakes contexts where fundamental human interests are at issue, this is insufficient on its own to discharge the agent's justificatory responsibility. This is so because the very purpose of retaining a human in the loop in contexts like these is precisely to ensure that automated recommendations are treated as defeasible inputs to deliberation rather than authoritative resolutions. Agents may ultimately decide to follow the system's output, but what is normatively decisive in this move is that they retain deliberative ownership of the decision through critical reflection on the action they endorse. Thus the failure to exercise independent judgment in these

circumstances represents not merely a breakdown in human–system interaction but a diminishment of the moral significance of keeping humans in the loop or on the loop in the first place. This is perhaps the most troubling implication of strong automation bias.

4 Cultivating Critical Trust Against Strong Automation Bias

Automation bias is a complex socio-technical phenomenon, and it should come as no surprise that scholars have proposed various, quite different mechanisms as means for its mitigation or amelioration. With respect to what we distinguished above as weak automation bias, Skitka and her colleagues have persuasively argued that emphasizing accountability prior to decision-making encourages more cognitively complex and self-critical information processing, which contributes to reducing the incidence of automation bias. Their study shows that accountability reduced both commission and omission errors, but only when participants were accountable for overall performance or accuracy; accountability for tracking or having no accountability led to higher error rates. [15] In recent years, there has been an increasing number of studies on the role that explanations can have in mitigating automation bias. The findings of some of them are that explanations do not systematically reduce the users' tendency to accept incorrect automation suggestions and that, in some cases, the explanations actually increased the bias (Vered et al., 2023). This seems to confirm earlier findings concerning trust in automation, which claimed that explanations tend to reinforce trust in systems (Dzindolet et al., 2002, 2003).

If our arguments about the conceptual features and normative significance of strong automation bias are correct, then one natural mitigation strategy is to implement design-based mechanisms that address its root cause: the excessive trust in system capabilities. As discussed above, this overtrust is sustained by the personal positive experience with autonomation, the reputation of automated systems, reinforced through widely circulated myths about the purported objectivity, neutrality, and infallibility of automated outputs, as well as the institutional reinforcement of their authority through different organizational policies and protocols. We believe, however, that in tasks that permit automation without stringent time pressures, these factors can be counteracted by introducing mechanisms that cultivate epistemic friction within socio-technical systems—that is, design features that re-engage users' deliberative capacities when responding to automated cues. Until the end of this paper we will discuss two such design tools for countering strong automation bias: Reflection Machines and defeaters. Both tools aim to mitigate the excessive trusts at the root of strong automation bias, but they do so through distinct mechanisms and with different scopes. Reflection Machines function as meta-cognitive prompts designed to sustain critical reflection in human users. Defeaters, by contrast, operate as explicit signals embedded within systems that alert users to potential unreliability or error demanding the operator to actively evaluate the output and make a decision by sampling other indicators. Before proceeding, it is important to mention that

[15] For skepticism about the mitigating effect of accountability mechanism see Tetlock & Boettger (1989).

although these interventions are justified here in relation to strong automation bias, their justification is not limited to this phenomenon. They can also be claimed to have an effect of reducing weak automation bias, or serve to enhance or preserve other values within socio-technical systems, such as for instance autonomy (Buijsman et al., 2025).

4.1 Reflection Machines

Reflection Machines or reflective agents in general are meta-cognitive tools that actively prompt operators to engage critically with automated outputs. Rather than allowing the cues of an automated system to be passively accepted, these tools encourage critical reflection, scrutiny, and the exercise of an independent human judgment, counteracting the tendency toward mere deference to the system. In doing this Reflection Machines seem to offer resources for both countering the negligence implied by weak automation bias as well as the deference implied by strong automation bias.

As Haselager et al. (2024) argue in the context of Reflection Machines developed for medical decision-support systems, these tools do not simply generate or optimize cues; instead, they act as meta-prompts on users to interrogate automated cues by asking why a particular cue was made, whether alternative interpretations exist, and which contextual factors may have been overlooked. For instance, a Reflection Machine may highlight data inconsistencies, draw attention to atypical patient profiles, or ask to consider less likely but clinically relevant alternatives. The integration of Reflection Machines in a particular decision-support system should be seen as a deliberate introduction of epistemic friction into the interaction between the human and the decision-support system and as a specific *normative* requirement. The purpose is to ensure that, through a mandatory engagement with a Reflection Machine, the operators are prompted to pause, actively reflect and (re)assess an automated cue before acting on it. This way Reflection Machines create opportunities for the operators to take "deliberative ownership" of their actions on a cue and ensure that they do not defer non-autonomously to the system.

A natural question emerges at this point.[16] What if the reflective agent itself fails to surface relevant concerns, frames the deliberation inadequately, or embeds its own biases? The very rationale for deploying reflective agents is to mitigate the risks of automation bias; yet isn't it the case that their introduction reproduces the same structural problem at a higher level? We believe that the risk of automation bias in reflective agents differs and is considerably more limited than that in other types of automation for two main reasons. First, reflective agents do not output facts. They are not intended to provide information or inferences that are simply correct or incorrect. Instead, reflective agents challenge humans to think critically about the issue raised in the conversation. The content of the agent's questions does not need to be strictly determined in any epistemic sense; rather, it must be appropriate and relevant to the conversational context. So, the reflective agent simply prompts operators to elaborate on the values underlying potentially morally sensitive decisions, thereby helping

[16] We are thankful to an anonymous reviewer for pressing this point.

to enhance the insight into the values at stake, their own position, and the alternative perspectives. For example, in context of systems for automated face recognition for public safety Grauwde et al. (2023 and Unpublished manuscript) apply agonistic deliberation and stage-based reflection—using follow-up questioning, agonistic inquiry, and breakdown methods— that enable constructive disagreement and deeper reflection.This approach allows users to examine ultimately how different value commitments might justify or challenge particular choices and to articulate the trade-offs and uncertainties inherent in their own position. Recent research suggests that integrating such reflection mechanisms into the use of AI technology can enhance users' critical capacities and thereby provide opportunities to mitigate automation bias (Wingerter et al., 2025).

A second reason why reflective machines appear to pose comparatively limited risks of automation bias concerns recent advances in techniques designed to constrain and stabilize large language model behavior. Contemporary approaches to prompt design, response validation, etc., aim to keep system outputs bounded within relevant conversational and informational contexts. These methods reduce the likelihood that the system will generate confabulatory[17] or misleading content that might inadvertently acquire unwarranted epistemic authority in the eyes of users. This matters because, as we argued above, automation bias is triggered by patterns of interaction in which system outputs are perceived as authoritative, reliable, or trustworthy. Systems that frequently produce confident but inaccurate responses risk encouraging precisely such misplaced trust. By contrast, reflective agents designed to operate within constrained conversational frames, and explicitly positioned as prompts for reflection rather than sources of factual determination, present fewer cues that invite uncritical deference. Even where confabulatory elements arise, their role within agonistic or exploratory dialogue may itself become an object of reflection rather than a basis for compliance (cf. Sui et al., 2024).

Two additional concerns regarding the integration of Reflection Machines into decision-support systems merit consideration. First, questions arise about their applicability across domains, particularly in environments characterized by severe time constraints. In contexts where decisions must be made within seconds, opportunities for structured reflection may be limited. This is a credible concern. We do not suggest that reflective agents can be universally deployed across all decision environments. Their implementation presupposes situations in which at least minimal discretionary time for reflective engagement is available.

Nevertheless, this limitation does not preclude their relevance in many high-risk domains. As discussed above, reflective agents are already being explored in medicine and public safety contexts, where reflection can be integrated into diagnostic, supervisory, or evaluative stages of decision-making. Similar strategies are observable in military settings. Team design patterns allocate moral and evaluative decision-making to specific phases of operations—for example, before-, in-, or on-the-loop stages in surveillance or targeting missions (van der Waa et al., 2020; van Diggelen et al.,

[17] In generative-AI communities this phenomenon is often described as "hallucinations," although this metaphor is somewhat misleading: hallucinations involve sensory experiences without corresponding external stimuli, whereas LLMs do not possess sensory experience (Smith et al., 2023).

2024). These allocations aim to ensure that the morally accountable or answerable actor possesses the authority, expertise, situational awareness, and temporal capacity required for judgment. Reflective agents can operate at these decision nodes by prompting scrutiny of assumptions, surfacing value tensions, or probing affective and evaluative responses relevant to morally charged choices (van Diggelen et al., 2023).

It is also important to situate Reflection Machines within the broader landscape of safeguards already present in high-risk socio-technical systems. Domains such as medicine, aviation, and the military employ multiple mechanisms to mitigate automation-related failure, including human-in/on-the-loop control, redundancy protocols, structured accountability arrangements, trust-calibration training, uncertainty displays, and explainability tools. These measures primarily address system reliability, error detection, and procedural responsibility. By contrast, Reflection Machines complement rather than replace existing safeguards. Within military contexts, including decision points along the kill chain, Reflection Machines should not be understood as an additional authorization layer. Their role is instead that of a cognitive safeguard embedded at critical decision nodes, designed to preserve evaluative engagement precisely where independent human judgment is normatively required.

A second worry is that although such tools may plausibly generate opportunities for critical reflection, their mere presence does not in itself guarantee deliberative ownership or the exercise of moral agency which we emphasized above. Moreover, one may worry that mandatory interaction with these systems may risk degenerating into procedural compliance, thereby reproducing the very concerns associated with "rubber-stamping," albeit at a more sophisticated level. In response, we must first emphasize that the mere presence of reflective agents is not assumed to automatically *secure* deliberative ownership and the exercise of moral agency. Rather, reflective systems are designed to *scaffold* conditions under which such ownership becomes more likely. So the normative force of the intervention does not lie in *guaranteeing* autonomous actions or the exercise of moral agency, but in reshaping the structure of human–system interaction so that deference becomes more difficult to sustain. The concern that mandatory interaction may degenerate into procedural compliance seems well-founded. Any institutionalized reflective practice risks routinization. However, this risk is not unique to reflective agents: it applies to many existing safeguards, including checklists, justification protocols, and human-in-the-loop authorization procedures. The relevant problem should therefore be considered as being design-sensitive. If properly designed reflective systems resist rubber-stamping by requiring context-sensitive elaboration, generating unpredictable follow-up prompts, and dynamically adapting to the substance rather than the form of user responses. There is a substantial body of research on methods for establishing and assessing high levels of engagement in human–agent interaction (Oertel et al., 2020), while large language models (LLMs) provide new opportunities to further advance such engagement through open-ended, adaptive dialogue. For example, recent work on LLM-based Socratic conversational agents demonstrates their potential to foster cognitive engagement and reflective thinking by structuring dialogue around probing and elaborative questioning (e.g., Xi et al., 2026). An important concern, however, remains in regard of the long-term use of such LLM-based agents. Over time, both the user and the LLM may mutually adapt their speech acts, interaction patterns, and

expectations. This co-adaptation may influence the depth, direction, and quality of critical reflection. Future research must therefore develop strategies to ensure that such adaptive processes evolve in desirable directions—for example, by sustaining epistemic challenge, preserving critical distance, and preventing convergence toward superficial or confirmatory interaction patterns.

4.2 Defeaters

Another design-based proposal for mitigation of strong automation bias can be found in the "Design for Defeaters" framework, proposed by Veluwenkamp and Buijsman (2025) which has been explicitly tailored for improving the authority and the epistemic position of the human operators in a way that will allow for contestation of the cues of the decision support system. The framework purports to achieve this aim through the integration of defeaters, understood as system-integrated mechanisms which highlight potential flaws in the decision-making process of the decision-support system. Veluwenkamp and Buijsman categorize the defeaters into two groups: *undercutting defeaters* which address the reliability of data based on which a cue was delivered and *rebutting defeaters* which directly contests an automated cue. These latter, we find to be more appropriate for countering automation bias as they can serve for highlighting the evidence that contradicts the automated cue or for offering alternative interpretations.

For example imagine that a medical decision-support system recommends diagnosing a patient with Condition A on the basis of imaging data. Unlike an undercutting defeater, which would challenge the reliability of the input data or the system's processing pipeline, a rebutting defeater introduces evidence that supports an incompatible diagnosis. When the system recommends Condition A, the rebutting defeater might automatically display clinical indicators from the patient record—such as lab values or symptom patterns—that statistically correlate with Condition B rather than Condition A; or it might surface peer-reviewed differential-diagnosis criteria showing that specific features visible in the scan are more predictive of Condition B. In each case, the defeater directly contests the system's recommendation by supplying counterevidence, inciting the clinician to exercise one's judgment on whether the automated cue should be accepted, qualified, or rejected.

One concern related to the defeaters framework is whether such arrangements might introduce the risk of cognitive overload, potentially burdening the operator with an excess of competing interpretations or evidential inputs. This worry seems warranted. It is important to acknowledge — as the literature on defeaters also notes — that introducing too many defeaters risks overcomplicating the decision environment. On the one hand, the accumulation of decision points could effectively shift the burden of judgment back onto the human operator, thereby diminishing the functional role of the automated support system. On the other hand, operators may respond by disregarding defeaters altogether, or by treating them as routine formalities, which would reintroduce the very disengagement from critical evaluation these mechanisms are meant to prevent. We do not take it to be necessary to resolve this tension within the present analysis. However, we agree that it merits explicit recognition as a practical and empirical challenge. Determining the appropriate balance between

reflective friction and cognitive manageability is likely to depend on contextual factors and therefore requires further domain-specific investigation.

Note that, similar to the case of Reflection Machines, the implementation of defeaters also introduces epistemic friction into the interaction between the human operator and the decision-support system. Practically, this design approach helps mitigate automation bias by making system limitations—whether in respect to the data used or to a particular cue generated by the system—more salient. To the extent that an operator's trust calibration is sensitive to the perception of these limitations, the framework also helps sustain critical trust within the socio-technical system. At the same time, the introduction of rebutting defeaters appears also to target weak automation bias. The salience of the defeaters invites operators to interpret these signals, weigh them against the AI's suggestion, compare alternatives, reason through trade-offs, maintain vigilance in critical moments, and decide whether to accept or contest a cue. By explicitly combining defeater signals with sufficient cognitive bandwidth, the framework enhances the capacity of human decision-makers to retain both the normative and epistemic authority necessary to counter strong automation bias and to maintain, more generally, critical and accountable human judgment.

5 Conclusion

Strong automation bias, understood as epistemic deference, poses distinctive ethical and moral challenges in human–automation interaction. When users defer to automated cues despite having access to contradictory evidence that supports an alternative course of action, they may compromise their own autonomous agency and fail to exercise their moral agency that high-stakes contexts demand. Addressing these challenges requires interventions that target the root cause of strong automation bias, namely the excessive trust in automation. Some design-based mechanisms—such as Reflection Machines and defeaters—may well serve this purpose. By prompting critical scrutiny of system outputs, highlighting contextual uncertainties, and presenting alternative interpretations, these mechanisms ensure that automated recommendations are treated as contributory inputs rather than binding directives. Together, both tools allow human decision-makers to benefit from automation while retaining the evaluative independence necessary for the quality of human judgment in high-stakes environments.

Funding None.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication All authors consent to submission and publication at Philosophy and Technology.

Competing interests The author(s) declare that there are no conflicts of interest.

References

Abraham, Y. (2024). 'Lavender': The AI machine directing Israel's bombing spree in Gaza. *+972 Magazine*. https://www.972mag.com/lavender-ai-israeli-army-gaza/

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*, 153–169.

Bokros, S. E. (2021). A deference model of epistemic authority. *Synthese, 198*(12), 12041–12069.

Brownstein, M., & Jennifer Saul, eds. (2016). Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology. Oxford: Oxford University Press.

Buijsman, S., Carter, S. E., & Bermúdez, J. P. (2025). Autonomy by Design: Preserving Human Autonomy in AI Decision-Support. *Philosophy & Technology*, *38*, 97.

Christman, J. (1991). Autonomy and personal history. *Canadian Journal of Philosophy, 21*(1), 1–24.

Coco, A. (2023). Exploring the impact of automation bias and complacency on individual criminal responsibility for war crimes. *Journal of International Criminal Justice, 21*(5), 1077–1096.

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. *AIAA 1st Intelligent Systems Technical Conference*.

Davia, C. (2015). Moral deference and deference to an epistemic peer. *Philosophical Quarterly, 65*(261), 605–625.

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.

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology, 17*(3), 155–163.

Downey, A. (2025). The alibi of AI: Algorithmic models of automated killing. *Digital War, 6*, Article 9.

Dratsch, T., Chen, X., Mehrizi, M. R., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Santos, DPd. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology, 307*(4), Article e222176.

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge University Press.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697–718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*, 79–94.

Feinberg, J. (1986). *The Moral Limits of the Criminal Law. Volume 3: Harm to Self.* Oxford University Press.

Fiske, S. T., & Taylor, S. E. (1994). *Social Cognition.* (2nd ed). McGraw-Hill.

Frankfurt, H. G. (1988). *The Importance of What We Care About: Philosophical Essays.* Cambridge University Press.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127.

Grauwde, M., Neerincx, M., & Kudina, O. (2023). Conversational agents for a deliberative age. In *Works-in-progress and demonstrations track*. The Eleventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP).

Grauwde, M., Neerincx, M., & Kudina, O. (Unpublished manuscript). *Scaffolding stakeholder reflection for value-aligned public-safety AI: A conversational agent to support deliberation*.

Gravett, W. H. (2023). Judicial decision-making in the age of artificial intelligence. In Henrique Sousa Antunes, Pedro Miguel Freitas, Arlindo L. Oliveira, Clara Martins Pereira, Elsa Vaz de Sequeira, Luís Barreto Xavier (Eds.), *Multidisciplinary perspectives on artificial intelligence and the law* (vol. 58, pp. 281–297). Law, Governance and Technology Series. Cham: Springer Nature.

Haselager, P., Schraffenberger, H., Thill, S., Fischer, S., Lanillos, P., Sebastiaan, van de Groes, & Miranda van Hooff. (2024). Reflection machines: Supporting effective human oversight over medical decision support systems. *Cambridge Quarterly of Healthcare Ethics, 33*(3), 380–389.

Heyns, C. (2017). Autonomous weapons in armed conflict and the right to a dignified life: An African perspective. *South African Journal on Human Rights, 33*(1), 46–71.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*, 407–434.

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43*(3), 274–306.

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese, 198*(10), 9941–9961.

Johnson, G. M. (2024). Varieties of bias. *Philosophy Compass, 19*(7), Article e13011.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). Judgment under Uncertainty: Heuristics and Biases. New York: Cambridge University Press.

Kelly, T. (2022). *Bias: A Philosophical Study*. Oxford University Press.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, Article 108352.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2019). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors, 62*, 123–142.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Levy, N. (2016). Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research, 93*(3), 3–26.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association, 24*(2), 423–431.

Lyell, D., Magrabi, F., Raban, M. Z., Pont, L. G., Baysari, M. T., Day, R. O., & Coiera, E. W. (2017). Automation bias in electronic prescribing. *BMC Medical Informatics and Decision Making, 17*(1), 17–28.

Mackenzie, C. (2008). Relational autonomy, normative authority and perfectionism. *Journal of Social Philosophy Vol, 39*(Winter), 512–533.

McKenna, M. (2005). The Relationship between Autonomous and Morally Responsible Agency. In J. S. Taylor (Ed.),

Mosier, K. L., Everett, A., Palmer, & Degani, A. (1992). Electronic checklists: Implications for decision making. *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 7–11).

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology, 8*(1), 47–63.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*, 527–539.

Natali, C., Marconi, L., Duran, L. D. D., & Cabitza, F. (2025). AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. *Artificial Intelligence Review, 58*, Article 356.

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., & Peters, C. (2020). Engagement in human–agent interaction: An overview. *Frontiers in Robotics and AI, 7*, Article 92.

Palmira, M. (2020). Expert deference about the epistemic and its metaepistemological significance. *Canadian Journal of Philosophy, 50*(4), 524–538.

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(2), 230–253.

Peters, U., Krauss, A., & Braganza, O. (2022). Generalization Bias in Science. *Cognitive Science*, *46*(9), e13188.

Pozzi, G. (2023). Automated opioid risk scores: A case for Machine Learning-Induced Epistemic Injustice in Healthcare. *Ethics and Information Technology, 25*, Article 3.

Romeo, G., & Conti, D. (2025). *Exploring automation bias in human–AI collaboration: A review and implications for explainable AI* (vol. 3July). AI & Society.

Rosbach, E., Ganz, J., Ammeling, J., Riener, A., & Aubreville, M. (2025). Automation bias in AI-assisted medical decision-making under time pressure in computational pathology. In Christoph Palm, Katharina Breininger, Thomas Deserno, Heinz Handels, Andreas Maier, Klaus H. Maier-Hein, Thomas M. Tolxodorff (Eds.), *Bildverarbeitung für die Medizin 2025* (pp. 129–134). Springer Fachmedien.

Saul, J. (2013). Scepticism and implicit bias. *Disputatio, 5*(37), 243–263.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard University Press.

Schemmer, M., Kühl, N., Benz, C., & Satzger, G. (2022). On the influence of explainable AI on automation bias. *Thirtieth European Conference on Information Systems (ECIS 2022), Timişoara, Romania.,1–12.*

Skitka, L. J., Mosier, K., Burdick M.D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52*(4), 701–717.

Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health, 2*(11),1–3.

Sparrow, R. (2016). Robots and respect: Assessing the case against Autonomous Weapon Systems. *Ethics & International Affairs, 30*(1), 93–116.

Stoljar, N. (2000). "Autonomy and the Feminist Intuition." In C. Mackenzie & N. Stoljar (Eds.), *Relational Autonomy: Feminist Perspectives on Autonomy, Agency and the Social Self* (pp. 94-111). New York/Oxford: Oxford University Press.

Sui, P., Duede, E., Wu, S., & So, R. (2024). Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (vol. 1, pp. 14274–14284). Long Papers.

Tetlock, P. E., & Boettger, R. (1989). Accountability: A Social Magnifier of the Dilution Effect. *Journal of Personality and Social Psychology*, *57*(3), 388–398.

Tilton, E., & Briana Toole. (2010). Standpoint epistemology and the epistemology of deference. In Mathias Steup (Ed.) *Blackwell companion to epistemology*, (3rd ed.) (pp. 584–594). Malden, MA: Blackwell.

Ullrich, D., Butz, A., & Diefenbach, S. (2021). The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction. *Frontiers in Robotics and AI*, *8*, 554578.

van der Waa, J., van Diggelen, J., Siebert, L. C., Neerincx, M., & Jonker, C. (2020). Allocation of moral decision-making in human–agent teams: A pattern approach. In *Human–computer interaction – HCI 2020* (pp. 203–220). Springer International Publishing.

van Diggelen, J., van den Bosch, K., Neerincx, M., & Steen, M. (2024). Designing for Meaningful Human Control in Military Human–Machine Teams. In Giulio Mecacci, Daniele Amoroso, Luciano Cavalcante Siebert, David Abbink, Jeroen van den Hoven and Filippo Santoni de Sio (Eds.), *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* (pp. 232–252). Edward Elgar Publishing.

van Diggelen, Jurriaan, Metcalfe, J.S., van der Bosch, K., Neerincx, M., Kerstholt, J.(2023). Role of Emotions in Responsible Military AI. *Ethics and Information Technology*, *25*(1), 17.

Veluwenkamp, H., & Buijsman, S. (2025). Design for operator contestability: Control over Autonomous Systems by Introducing Defeaters. *AI Ethics, 5*, 3699–3711.

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence, 322*, Article 103952, 1–24.

Wingerter, T., Lewis, T., Straub, & Schweitzer, S. (2025). Mitigating Automation Bias in Generative AI through Nudges: A Cognitive Reflection Test Study. *Procedia Computer Science*, *270*, 2106–2114.

Xi, L., Zhang, Y., & Wang, Q. (2026). Investigating the effects of an LLM-based socratic conversational agent on students' academic performance and reflective thinking in higher education. *Computers & Education*, 241, 105494.

Zerilli, J.,Goñi, I., & Placci, M.M. (2024). Automation bias and procedural fairness: A short guide for the UK civil service. BRAID Reports, 1–22.