
Trust Hinders Bidirectional Human–AI Alignment

Michaela Benk

Department of Informatics
University of Zurich, Switzerland
michaela.benk@uzh.ch

Kacper Sokol

Department of Informatics
USI Lugano, Switzerland
kacper.sokol@usi.ch

Abstract

Trust calibration remains the dominant framing of AI-assisted decision making: the goal is to close the gap between perceived and actual trustworthiness of AI systems so that users exhibit appropriate reliance on them. We argue that this approach, while useful for decision problems that are well-defined and clearly scoped, overly restricts what it means to make a decision with AI support. To overcome this limitation, we develop three dimensions of AI-assisted decision making. First, trusting AI is only one of several possible strategies for managing uncertainty. Second, trust calibration is often misapplied when AI models “large-world” decision problems, i.e., ill-defined situations where goals and states are only partially known or unknowable. Third, in such settings human–AI alignment requires addressing higher-order decision processes. We conclude by reflecting on the implications of our observations on the design of process-oriented human–AI coupled systems that jointly identify and negotiate uncertainty throughout the decision-making task.

1 Introduction

Imagine you are hiking across an unfamiliar terrain when you reach a fork in the trail. Heavy fog obscures both paths; a thunderstorm is approaching. Conflicting reasons come to mind in favour and against each route. In other words: *you are uncertain*. You ask yourself which path is faster. You also wonder whether choosing immediately is the right strategy. Maybe you should rather wait for the fog to clear, seek higher ground, turn back or reach out to a more experienced hiker for advice.

Suppose you instead consult a decision support tool (DST) that uses an artificial intelligence (AI) model trained on relevant hiking data. It recommends the left path, which is shorter, hence should take less time. Assume the suggestion is correct. You follow it, but the path lacks any trees to shelter you from the rain. The storm intensifies, you slip on wet ground and suffer a head injury.

What, if anything, does your trusting the DST tell us about the efficacy of your decision? Was your choice good because the underlying recommendation was trustworthy? Or was your decision bad because of the resulting head injury? Would your partner or the medical professional treating you answer this question in the same way as the designer of the DST you used given the outcome?

Despite the richness of such decision scenarios, much of the literature on AI-assisted decision making evaluates this process primarily in terms of achieving well-calibrated trust and appropriate reliance [20, 21, 14, 13]. Yet in the vignette introduced above, the decision quality does not primarily depend on whether the AI model was trustworthy or the hiker exactly followed its advice. Rather, it is determined by the joint alignment spanning multiple levels of uncertainty that the decision maker faces as well as the appropriate strategy to navigate it, which may or may not be to uncritically select a path that is recommended as the shortest.

The scenario described throughout this introduction aims to motivate our argument: focusing primarily on trust confines AI assistance to an overly narrow conception of (human) decision making, thus hindering the advancement of bidirectional human–AI alignment. To substantiate this claim, we first overview relevant concepts (Section 2) and then develop three dimensions of AI-assisted decision making. These span: (i) strategies to manage uncertainty (Section 3), (ii) characteristics of the decision problem (Section 4) and (iii) levels of higher-order decision processes (Section 5). Specifically, we show that trust calibration is only one of several possible approaches for managing uncertainty in decision making. We argue that this particular strategy is useful for a confined class of problems, namely those characterised as “small-world”, where the decision space and relevant states are well-defined and enumerable [6]. In “large-world” problems, on the other hand, uncertainty may concern the decision space itself as well as high-level objectives, availability of alternatives and best strategies for their implementation. In such settings, humans and AI tools may hold fundamentally different framings of the decision problem. To achieve bidirectional human–AI alignment, we therefore propose shifting attention from trust calibration and reliance to understanding how humans and AI can jointly uncover, navigate and negotiate the uncertainty spanning different levels of AI assistance. We finish with a discussion of limitations and counterarguments (Section 6) before concluding (Section 7).

2 Preliminaries and Definitions

Our three dimensions of AI-assisted decision making are built upon the notions of “small-” and “large-world” decision problems (Section 2.1) and trust calibration (Section 2.2).

2.1 Flavours of Uncertainty in “Small-” and “Large-world” Decision Problems

Decision theorists distinguish between situations of **risk**, **ambiguity** and **uncertainty**, accounting for how they affect information provision [16]. The former two exist within “small-world” problems, whereas the latter one characterises “large-world” problems [6].

Formally, in a situation of **risk**, the set of all the future states together with the set of all their consequences are known. Taking a risk – or a chance – is essentially a form of a gamble. In contrast, when the probabilities of the consequences are unknown, but in principle knowable, it becomes a situation of **ambiguity**. Further insights can thus close the information gap when dealing with both risk and ambiguity.

Uncertainty, on the other hand, refers to ill-defined settings. Here, the future states and consequences – or even the full account of the current situation – are neither known nor knowable because this information cannot be obtained or changes at random. Crucially, the Bayesian theory of expected utility maximisation applies *only* to “small-world” problems. Savage [16] went as far as considering it “ridiculous” to apply it in situations of uncertainty.

We return to how these different flavours of uncertainty apply to AI-assisted decision making in Section 4.

2.2 Origins, Definitions and Measurements of Trust Calibration

The idea that trust can be calibrated to the right degree, such that appropriate reliance follows, originally arose in the trust-in-automation literature [1, 11, 5]. The central design problem therein was to help operators of automated systems deployed in work or industrial settings to decide when to rely on their tools. In this context, trust in an automated system is considered appropriate, calibrated or warranted when it is *caused* by the trustworthiness of the system [9].

Although trust calibration and appropriate reliance are conceptually distinct, they are operationalised in a largely the same way [1, 11, 5]. For example, the number of times that a user agrees with and decides to follow AI predictions has been linked to both of these concepts [23, 22, 2, 4]. A typical way to test whether trust in an AI tool is appropriate thus boils down to manipulating its trustworthiness and measuring the level of reliance exhibited by its users [9].

3 Trusting AI Is One of Many Strategies for Managing Uncertainty

Miller’s [12] discussion of different reasoning processes involved in AI-assisted decision making hints at a crucial, yet frequently left implicit, distinction: decisions come in *layers*. A *first-order* decision concerns the choice itself; for example, given a fork in a trail, which path should I take? A *higher-order* decision concerns how the first-order choice is made; for example, which strategy, rule or source of advice should I use when deciding?

Sunstein and Ullmann-Margalit [19] provide an overview of key strategies people adopt at a higher order: rules, heuristics, presumptions, standards, routines, small steps, picking and delegating (see Table 1 in Appendix A for their description). Several of these appear in the introductory vignette. Such strategies are often chosen *before* the first-order situation arises, precisely to reduce the cognitive burden of any eventual choice. They work by restricting what the decision maker has to consider. Deciding whether to act on external advice, like an AI recommendation, is, in Sunstein and Ullmann-Margalit’s [19] terms, a second-order *delegation* decision. Consequently, trust calibration sits at a meta level: it evaluates whether the delegate deserves the delegation. Yet the literature mostly treats this higher-order concern as the primary decision problem, obscuring the process that characterises real-world decisions and hindering what human–AI alignment requires in such cases.

4 Trust Calibration Is Misapplied to “Large-world” Decision Problems

Since its origins in the trust-in-automation literature, the trust calibration paradigm has been extended well beyond its intended setting, now spanning medical diagnosis, law enforcement and financial investment [23]; most recently, it was applied to large language models supporting decisions across personal, legal, medical and scientific domains [14]. What makes this problematic is that the underlying assumptions were adopted uncritically in these new contexts, where their suitability has not been examined or validated.

Trust calibration is defined as the correspondence between perceived and actual trustworthiness [9], which presupposes access to an oracle judging *correctness* of AI outputs. In the introductory vignette, the ground truth pertaining to which path is the shortest, is known. Yet this type of information remains unknowable for most “large-world” decisions [18]. Returning to our hiking scenario, asking which path is the fastest or which one will not result in a head injury is ill-specified or unanswerable.

Trust calibration presupposes a *stable* reference point for assessing trustworthiness; in the hiking vignette, the measured distance of a path is unlikely to change. In human–AI coupled systems, however, the AI may alter a user’s beliefs, preferences or goals, thereby shifting the very benchmark against which its trustworthiness is judged. As decision-aid scholars have long noted, preferences are rarely well-formed; they rather emerge and evolve throughout the decision process itself [15]. Unsurprisingly, empirical studies tend to circumvent this limitation by imposing artificial stability.

Moreover, trust calibration presumes that the decision frame is *fixed*; relevant states, consequences, objectives and alternatives are all known. In this setting, trust calibration “flattens” the decision process to a single question: whether to trust or not. This narrow focus can nonetheless mask fundamental misalignments on other levels. For example, if the hiker in our vignette follows a trustworthy AI recommendation and gets hurt, nothing in the interaction revealed whether the choices and strategies considered were appropriate.

Although trust is generally recognised as a dynamic construct [3], trust calibration is measured through reliance on isolated recommendations, with its success evaluated by assessing joint performance. This focus on *outcomes* nonetheless obscures where the real difficulty lies when dealing with “large-world” decisions. As highlighted by Keren and De Bruin [10], once the structure and states of a problem are given, the decision itself is nearly trivial. This is because the difficulty resides in structuring the problem space as well as reflecting on the possible states and the uncertainties or probabilities associated with different outcomes.

Consequently, many empirical studies inadvertently reformulate the “large-world” problems that they attempt to address as to satisfy methodological assumptions, effectively creating ill-formed “small-world” problems.

5 Human–AI Alignment in “Large-world” Settings Requires Addressing Higher-order Decision Processes

We posit that addressing the challenges discussed thus far and achieving bidirectional human–AI alignment in decision making requires a shift in perspective. From the decision maker’s standpoint, delegation to a DST is one higher-order strategy among many. Choosing it presupposes that the decision problem (“small-” or “large-world”) as well as the structure and role of AI assistance are fixed; as meaningful interaction becomes redundant in this setting, bidirectional alignment opportunities vanish. We thus propose a rudimentary hierarchy that distinguishes several possible levels of AI assistance.

First-order: choice. What is the best choice among the available options?

Second-order: strategy selection. Which decision strategies are available and appropriate?

Third-order: decision structure. What are the relevant goals, objectives, alternatives and criteria?

At the most basic level, a decision maker chooses among a set of alternatives. When a DST proposes a single output and the user only decides whether to accept or reject it, the interaction follows the classical supervisory control paradigm from trust-in-automation research [11]. In this setting, the AI effectively makes the choice, and the human’s role is reduced to approving or discarding it. Trust calibration operates one tier above this choice as a proxy, which helps the user to determine whether to rely on the AI support or not in the first place. AI support can also enter at higher levels in a collaborative manner: it can help to structure the decision, identify appropriate decision strategies and determine whether delegation to AI is appropriate. Defining these levels enables us to situate trust calibration, as it is typically operationalised, within a broader hierarchy of decision-making processes. It further allows us to identify at what level uncertainty enters the picture and opens up a new space of interaction designs. We leave designing interfaces that allow humans and AI to negotiate types and objects of uncertainty for future work.

6 Limitations and Counterarguments

Moving beyond outcome-based evaluation is inherently difficult. A process-oriented approach judges decision quality by how well the decision structure represents the decision maker’s goals and whether an appropriate strategy was chosen [10]. Yet calling the hiker’s choice “successful” after sustaining a head injury may strike most people as absurd, regardless of whether the underlying process was sound. This psychological pull towards outcome evaluation is thus an obstacle to the shift we propose.

Another critique of our proposal is that trust calibration may simply require a better formalisation. Guo et al. [7], for instance, ground reliance in statistical decision theory, assuming known utilities, rational Bayesian updating and an optimal benchmark. But these are precisely the “small-world” assumptions we argue against: their framework works well because, and only when, the decision problem is already settled. Hullman et al. [8] support our view, showing that decision problems in AI-assisted decision-making studies are often underspecified. However, by suggesting that studies not adopting a utility-theoretic framework cannot provide sufficiently rigorous normative claims and should therefore be regarded as primarily descriptive, their conclusion anchors the field to a particular conception of decision making that rules out precisely the settings of most interest.

Crucially, our work criticises the use of trust calibration as a general-purpose paradigm in AI-assisted decision making. We neither argue that trust is irrelevant nor that it is the only focus of the field. Our claim is that the field uses it beyond its scope without examining whether its assumptions still hold.

7 Conclusion

AI-assisted decision making has been primarily positioned as a problem of trust calibration, allowing to identify when a DST is trustworthy and when people should rely on its recommendations. In this paper we argued that this framing is valuable for well-specified, “small-world” decision problems, but becomes limiting when applied to “large-world” settings that increasingly characterise the use of AI in practice. Rather than treating the calibration of human trust in a DST as a form of alignment, we situate it within a process in which humans and AI tools jointly uncover, contest and negotiate the

uncertainties that shape different levels of decision making. We contend that these considerations are especially important in view of the recent shift toward developing DSTs for critical thinking and reasoning support [12, 17, 18].

References

- [1] Michaela Benk, Sophie Kerstan, Florian Von Wangenheim, and Andrea Ferrario. 2025. Twenty-Four Years of Empirical Research on Trust in AI: A Bibliometric Review of Trends, Overlooked Issues, and Future Directions. *AI & SOCIETY* 40, 4 (April 2025), 2083–2106. doi:10.1007/s00146-024-02059-y
- [2] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–23. doi:10.1145/3706598.3714097
- [3] Cristiano Castelfranchi and Rino Falcone. 2010. *Trust Theory: A Socio-Cognitive and Computational Model* (1 ed.). Wiley. doi:10.1002/9780470519851
- [4] Rex Chen, Ruiyi Wang, Norman Sadeh, and Fei Fang. 2025. Missing Pieces: How Do Designs That Expose Uncertainty Longitudinally Impact Trust in AI Decision Aids? An In Situ Study of Gig Drivers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens Greece, 790–816. doi:10.1145/3715275.3732050
- [5] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies* 58, 6 (June 2003), 697–718. doi:10.1016/S1071-5819(03)00038-7
- [6] Gerd Gigerenzer. 2024. Psychological AI: Designing Algorithms Informed by Human Psychology. *Perspectives on Psychological Science* 19, 5 (Sept. 2024), 839–848. doi:10.1177/17456916231180597
- [7] Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica Hullman. 2024. A Decision Theoretic Framework for Measuring AI Reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 221–236. doi:10.1145/3630106.3658901
- [8] Jessica Hullman, Alex Kale, and Jason Hartline. 2025. Underspecified Human Decision Experiments Considered Harmful. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. doi:10.1145/3706598.3714063
- [9] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 624–635. doi:10.1145/3442188.3445923
- [10] Gideon Keren and Wändi Bruine De Bruin. 2003. On the Assessment of Decision Quality: Considerations Regarding Utility, Conflict and Accountability. In *Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making* (1 ed.), David Hardman and Laura Macchi (Eds.). Wiley, 347–363. doi:10.1002/047001332X.ch16
- [11] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan. 2004), 50–80. doi:10.1518/hfes.46.1.50_30392
- [12] Tim Miller. 2023. Explainable AI Is Dead, Long Live Explainable AI!: Hypothesis-Driven Decision Support Using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago, IL, USA, 333–342. doi:10.1145/3593013.3594001

- [13] Martin H. Möhle, Manuel Zall, Tim Thorwart-Gumpert, Nico Elbert, Frédéric Thiesse, and Christoph M. Flath. 2026. “Trust Me, Bro!” On the Influence of Agentic Support on Trust Calibration in AI-Assisted Decision Making. In *Artificial Intelligence in HCI*, Helmut Degen and Stavroula Ntoa (Eds.). Vol. 16744. Springer Nature Switzerland, Cham, 472–488. doi:10.1007/978-3-032-30860-3_30
- [14] Victor Ojewale, Julia Ryan, Suresh Venkatasubramanian, and C. Malik Boykin. 2026. More Is Not Better: Visual Uncertainty Cues and the Fragility of Trust Calibration in LLM-Assisted Decision Making. *Computers in Human Behavior: Artificial Humans* 8 (May 2026), 100307. doi:10.1016/j.chbah.2026.100307
- [15] Bernard Roy and Daniel Vanderpooten. 1996. The European School of MCDA: Emergence, Basic Features and Current Works. *Journal of Multi-Criteria Decision Analysis* 5, 1 (March 1996), 22–38. doi:10.1002/(SICI)1099-1360(199603)5:1<22::AID-MCDA93>3.0.CO;2-F
- [16] Leonard J. Savage. 1972. *The Foundations of Statistics* (2nd rev. ed.). Dover publ, New York.
- [17] Hua Shen, Tiffany Kneare, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2025. Position: Towards Bidirectional Human-AI Alignment. <http://arxiv.org/abs/2406.09264>. arXiv:2406.09264 [cs.HC] doi:10.48550/arXiv.2406.09264
- [18] Kacper Sokol, James Fackler, and Julia E Vogt. 2025. Artificial Intelligence Should Genuinely Support Clinical Reasoning and Decision Making to Bridge the Translational Gap. *npj Digital Medicine* 8, 1 (2025), 345. doi:10.1038/s41746-025-01725-9
- [19] Cass R. Sunstein and Edna Ullmann-Margalit. 1999. Second-Order Decisions. *Ethics* 110, 1 (Oct. 1999), 5–31. doi:10.1086/233202
- [20] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–39. doi:10.1145/3476068
- [21] Yingying Wang, Qin Ni, Tingjiang Wei, Haoxin Xu, Lu Liu, and Liang He. 2026. How AI-Assisted Decision-Making Paradigms and Explainability Shape Human-AI Collaboration. *Sustainability* 18, 7 (April 2026), 3516. doi:10.3390/su18073516
- [22] Yang Wu, Yao Wan, Mennatallah El-Assady, and April Yi Wang. 2026. StepMIND: A Visual Framework for Stepwise, Multimodal, and Bidirectional Explanations of AI-Generated Data Analysis Pipeline. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*. ACM, Paphos Cyprus, 522–539. doi:10.1145/3742413.3789070
- [23] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 295–305. doi:10.1145/3351095.3372852

A Overview of Second-order Strategies

Table 1 provides an overview of second-order strategies that people adopt to manage uncertainty.

Table 1: Second-order strategies for managing uncertainty according to Sunstein and Ullmann-Margalit [19]. Note that relying on an AI recommendation is an instance of *delegating* – see the last entry in the table.

Strategy	Description
Rules	Adopt a rule in anticipation of challenging or repetitive decisions; for example, “when it rains, always take the fastest route”. A rule is a full, or nearly full, <i>ex ante</i> specification of results in individual cases. Rules are error-prone in a characteristic way: they are over- and under-inclusive relative to the reasons that justify them.
Heuristics	Rely on a mental shortcut that bypasses case-by-case choices.
Presumptions	Act on the assumption that P is true until circumstances A arise, in which case stop relying on P and reconsider the situation.
Standards	Adopt a vague, value-based criterion that structures judgment without settling it; for example, drive “responsibly”, act “reasonably” and convict only “beyond a reasonable doubt”.
Routines	Rely on a behavioural pattern that is habit-like but more voluntary and self-conscious than a habit; for example, lock the office door on every exit or take the same commute daily. Occasional error is accepted as the price of not deciding.
Small steps	Make a narrow, reversible decision now and defer answering the larger question.
Picking	Decide on an arbitrary or random basis rather than choosing for a reason; for example, flip a coin, letting an irrelevant factor be decisive.
Delegating	Hand the decision to another party, like a spouse, a colleague or an AI tool. Delegation ranges from mere consultation all the way to fully relinquished control.