Computers in Human Behavior: Artificial Humans 6 (2025) 100216

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

Investigating choices regarding the accuracy-transparency trade-off of AI-based systems across contexts

Tim Hunsicker [a,*], Cornelius J. König [a], Markus Langer [b]

[a] Fachrichtung Psychologie, Universität des Saarlandes, Germany
[b] Institut für Psychologie, Universität Freiburg, Germany

ARTICLE INFO

Keywords:
Accuracy-transparency-trade-off
System choice
Artificial intelligence
Algorithmic decision-making
Trustworthiness
Framing effects
Deployer

ABSTRACT

Artificial intelligence (AI) is increasingly used in decision-making. However, choosing between different algorithmic methods underlying AI-based systems involves trade-offs. The accuracy-transparency trade-off is one of the most prominent: the most accurate approaches are often the least transparent, and the most transparent ones are the least accurate. This study examined how individuals navigate this trade-off from a deployer perspective. In an experimental between-participants online study ($N = 468$), we examined how framing (framing the system performance as accuracy rate vs. error rate), accountability (being able to justify a decision vs. no need to justify a decision), and the context of use (medicine, hiring, finance, law) affect choosing between different versions of systems underlying the accuracy-transparency trade-off. We also investigated whether the experimental manipulations and system choice affected trustworthiness and trust perceptions. Regarding the system choice (i.e., a preference for accuracy at the expense of transparency or a preference for transparency at the expense of accuracy), framing and accountability did not affect system choice. As expected, participants favored high performance in medicine compared to the other contexts. The results also supported the expected relationship between system choice and perceptions of different system trustworthiness facets, as well as framing effects on perceived trustworthiness and trust. We conclude that the context of use is critical for deployer preferences regarding system accuracy and transparency. Additionally, we identified person-related factors influencing such choices. Furthermore, a simple change in wording (i.e., without changing the system properties) can affect individuals' perceived trustworthiness of AI-based systems.

1. Introduction

As AI-based systems increasingly influence our lives, their properties and the ways we evaluate them are becoming ever more significant for a wide range of stakeholders, including providers, deployers, and policymakers. This is particularly pertinent in consequential decision-making contexts, such as medicine and hiring. Under the European AI Act, these contexts are classified as high-risk, imposing specific requirements (Section 2), such as technical documentation, transparency, provision of information to deployers, accuracy, robustness, and cybersecurity. These requirements place obligations on providers and deployers to ensure compliance (Section 3 of the AI Act). Among these, *system accuracy*—to minimize errors—and *system transparency*—to enhance control and understanding—are especially critical in high-risk settings (Adadi & Berrada, 2018). However, accuracy and transparency often conflict. The most accurate methods are frequently the least transparent and, therefore, the least understandable, while the most transparent methods tend to sacrifice accuracy, illustrating the accuracy-transparency trade-off (e.g., Assis et al., 2023). Navigating this trade-off is a central challenge for the responsible development and governance of AI.

This trade-off is particularly evident for systems that rely on highly accurate yet complex deep learning methods, which provide little insight into the decision-making process of the respective system (Burrell, 2016). In contrast, alternative approaches, such as rule-based methods, are more interpretable but often come at the cost of lower predictive accuracy (Rosenfeld & Richardson, 2019). Indeed, deploying such systems in real-world settings presents numerous challenges (Paleyes et al., 2023). Deployers of AI-based systems frequently face the challenge of choosing between these options, weighing the relative

* Corresponding author. Universität des Saarlandes, Campus A1 3, 66123, Saarbrücken, Germany.
E-mail addresses: tim.hunsicker@uni-saarland.de (T. Hunsicker), cornelius.koenig@uni-saarland.de (C.J. König), markus.langer@psychologie.uni-freiburg.de (M. Langer).

https://doi.org/10.1016/j.chbah.2025.100216
Received 12 August 2025; Received in revised form 1 October 2025; Accepted 1 October 2025
Available online 4 October 2025

importance of accuracy and transparency for a specific use case. This decision is not merely a technical calculation but a complex socio-technical judgment that has significant downstream consequences (Hatherley et al., 2024). Both transparency and accuracy are crucial for building trustworthy AI systems, particularly in high-risk domains, underscoring the need for a human-centered approach to design and development (Schoenherr et al., 2023). Adding to this complexity, the choice between different AI-based systems may also be influenced by how the properties of AI-based systems are described and framed (e.g., in system manuals). Providers must carefully determine how to communicate key attributes—such as accuracy and transparency—to stakeholders, including deployers, auditing institutions, and end-users (Landers & Behrend, 2023). For instance, framing these properties positively (e.g., the system is accurate in 95 % of cases) versus negatively (e.g., the system makes errors in 5 % of cases) can shape stakeholder perceptions, expectations, and ultimately, their choices between AI systems.

Research on the accuracy-transparency trade-off has focused on mapping where various algorithmic approaches fall within this spectrum (e.g., Hacker et al., 2020), attempting to quantify it (Bell et al., 2022), and highlighting the persistent nature of this trade-off (e.g., Adadi & Berrada, 2018; Assis et al., 2023). Meanwhile, research on explainable AI (XAI) has developed explainability approaches attempting to achieve understandability of highly accurate but intransparent systems (Guidotti et al., 2019; Miller, 2019). However, these explainability approaches also do not fully resolve the accuracy-transparency tradeoff. For example, they require validation themselves and may not always provide generalizable insights into system decision-making processes (Ali et al., 2023; Vilone & Longo, 2021). These challenges can be mitigated by opting for transparent algorithmic approaches, which are inherently more understandable and where explainability techniques can be more effective in enhancing insights (Rudin, 2019).[1] However, while these technical approaches are vital, they often overlook the subjective human factors that ultimately govern the selection of AI systems in practice.

Beyond the technical aspects of the accuracy-transparency trade-off, there is research that addresses the psychological implications of this trade-off (Dong et al., 2024; Nussberger et al., 2022). For instance, preferences for accuracy or transparency may vary depending on the context of use (Woodruff et al., 2020). However, there is limited evidence regarding which specific context of use leads to a preference for highly accurate or highly transparent algorithmic approaches. Furthermore, the properties of a (chosen) system, how these properties are framed and communicated (e.g., in system manuals), and the decision-maker's role when choosing a system may influence people's assessments of respective systems, especially the perceived trustworthiness (Hoff & Bashir, 2015; Schlicker & Langer, 2021). Such assessments in turn, shape the choice between systems, as well as future expectations and behaviors toward the system. We need to empirically understand how these situational factors, along with traits of the decision-maker, influence these choices along the accuracy-transparency trade-off.

This study aims to explore how individuals decide between different systems within the accuracy-transparency trade-off from the crucial perspective of the system deployer. We argue that this deployer perspective is essential for understanding real-world AI adoption, as deployers often act as organizational gatekeepers whose choices precede any end-user interaction and determine which systems get implemented, thus directly influencing downstream outcomes. We experimentally investigate whether the framing of properties of AI-based systems (e.g., system performance framed as accuracy vs. error rate), accountability for selecting a particular system, and the context of use (e.g., medicine vs. hiring) influence the choice between systems underlying the accuracy-transparency trade-off. The unique contribution of this paper lies in its combined investigation of these situational and personal factors across four distinct high-stakes domains. By doing so, our work provides comparative insights that move beyond prior research, which has often focused on a single domain or a more limited set of factors. Additionally, we assess whether these factors and the choice of system itself impact perceived trust and trustworthiness assessment of the respective system.

2. Background and hypotheses development

2.1. The accuracy-transparency trade-off

The choice of an algorithmic approach (e.g., manual or rule-based programming, based on machine learning or deep learning) as the basis for an AI-based system is accompanied by the decision of whether accurate or transparent processes and results are more important. In some cases, involved stakeholders may prioritize accuracy because it can result in improved decision-making or can ensure safety (EU Agency for fundamental rights, 2019). In contrast, they may prefer transparency when there is a strong desire for control of AI-based systems, which is a core aspect of human-centered system design, because transparency may allow for the explanation and justification of system outputs (Loi et al., 2021).

When considering differences in terms of these properties, there seems to be an inherent tension between accuracy and transparency (Gunning et al., 2019). Due to their complexity, the most accurate methods (e.g., deep learning based), without employing explainability approaches that try to generate post-hoc explanations, may offer no transparency about how decisions are reached, similar to a black box (e.g., Adadi & Berrada, 2018; Assis et al., 2023). The concept of transparency itself is multifaceted and can refer to the understandability of the entire model (simulatability), its individual components (decomposability), or its underlying algorithm (algorithmic transparency). In contrast, methods such as decision trees may be inherently transparent (Rosenfeld & Richardson, 2019). However, choosing more transparent methods usually compromises the accuracy of decisions, as they are less able to capture complex dependencies between input and output variables as, for example, deep learning methods (Arrieta et al., 2020). Eventually, choosing more transparent methods means that decisions can be better understood and better controlled (e.g., scrutinized, accepted/rejected based on an informed decision-making process) but simultaneously this increases the probability of less accurate predictions and erroneous system outputs. This establishes a psychological dilemma for the human decision-maker tasked with selecting a system.

2.2. Factors influencing the choice between different systems

When deployers of AI-based systems are confronted with the choice between different systems, they will be affected by the properties of the respective systems and their assessment of these properties. According to the European AI Act a deployer is "a natural or legal person, public authority, agency or other body using an AI system under its authority (…)" (Article 3). The deployer may be different from the provider, who "develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market (…)" (Article 3). The choice of an appropriate algorithmic approach thus is an important first step for the deployer in the implementation of an AI-based system for their context of use, followed by complying with the obligations. Deployers have to assess the

[1] Rather than adopting an absolute stance regarding the trade-off (seeing it as a fixed fact or denying it completely), we acknowledge the influence of contextual factors (e.g., resources or domain) which means that we consider the accuracy-transparency trade-off from a nuanced perspective (Crook et al., 2023).

information given by the providers of the systems and use them for their choice between different available AI-based systems in their specific context of use. In this study, we examine how four crucial aspects affect choices between different systems: the accuracy and transparency of the systems, how the accuracy is framed, the perceived accountability of the person making the decision, and the context of use. These factors represent a combination of system-level properties (accuracy, transparency), and situational variables (framing, accountability, contest) that influence the psychological processes of the decision-maker.

2.2.1. Accuracy and transparency – facets of system trustworthiness

Optimally, AI-based systems would be highly accurate *and* transparent. In fact, research on trust in automated systems implies that both system accuracy and transparency can be considered facets of system trustworthiness (Körber, 2019), and system trustworthiness should affect trust in system outputs as well as the choice for or against using a system (e.g., Hoff & Bashir, 2015; Schlicker & Langer, 2021).

Perceived trustworthiness consists of various facets that form the basis for trust (Lee & See, 2004; Mayer et al., 1995). For example, in their integrative model of interpersonal trust Mayer et al. (1995) proposed that perceived trustworthiness is based on the assessment of three characteristics of the trustee: *Ability* (domain-specific skills and competencies), *Benevolence* (the extent to which the trustee is believed to want to do good to the trustor), and *Integrity* (the perception that the trustee adheres to a set of principles that the trustor finds acceptable).

Models of trust in automation have adapted these concepts. For instance, Lee and See (2004) proposed three similar properties of a system: *Performance* (how well the system performs its function, parallel to Ability), *Process* (the principles and logic governing its operation, parallel to Integrity), and *Purpose* (the fundamental goals the system was designed to achieve and whether they align with the user's interests, which relates to Benevolence).

Both accuracy and transparency are crucial for the trustworthiness of AI systems, particularly in high-risk domains (Schoenherr et al., 2023). Whereas accuracy maps clearly to ability/performance, transparency does relate to integrity/process, but does not map perfectly onto a single facet in these earlier models. This is why the present study adopts the framework by Körber (2019), which integrates the models by Mayer et al. (1995) and Lee and See (2004) and contextualizes it in the context of advanced automation and AI. Crucially, Körber (2019) proposes that understandability is a critical aspect of perceived trustworthiness that becomes particularly salient with technologies (such as advanced machine learning) that lack transparency. In this study, we focus on two facets from that framework: perceived competence (or performance), which is associated with system ability or accuracy, and perceived understandability, which is associated with the transparency of system processes and outputs.

In reality, it may not be possible to have a highly accurate and highly transparent system. Thus, not only system developers must decide which property of a system to optimize, also stakeholders choosing between different systems (e.g., a company choosing between different AI-based systems for hiring or a hospital for diagnosing a disease) may be confronted with the fact that they cannot have a highly accurate and highly transparent system. They must decide what is more important in their respective context of use.

As a first step in the current study, we examine whether the preference of system choice (i.e., for transparency or accuracy) will also be reflected in participants' perceptions of facets of system trustworthiness. If participants favor system transparency at the expense of accuracy, the main focus of their choice may be on the traceability of processes and decisions, leading to a stronger perceived understandability of the system (Shin, 2021). If participants favor system accuracy, the main focus of their choice may be on increasing the number of correct system outputs, which should be reflected in a higher perceived system competence (Lee & See, 2004).

Hypothesis 1a. A system choice with a preference for transparency is associated with a higher degree of perceived understandability.[2]

Hypothesis 1b. A system choice with a preference for accuracy is associated with a higher degree of perceived competence.

2.2.2. Framing effects

The term *framing* refers to the fact that people can react differently to alternative descriptions of the same facts (Tversky & Kahneman, 1981). This concept is grounded in Prospect Theory, which was developed to describe human decision-making under risk. The theory distinguishes between two phases of the choice process: an initial "editing phase", where alternatives are analyzed and a subjective reference point is established, and a subsequent "evaluation phase", where those alternatives are assessed based on a subjective value function. Framing effects particularly affect the editing phase, as the description of a problem can shift the reference point from which outcomes are judged as gains or losses. For instance, for the same object, a different description or a change in perspective can alter assessments of and behaviors towards this object (Levin et al., 1998).

Levin et al. (1998) propose to distinguish between three types of framing effects: risky-choice, attribute, and goal framing effects. Of these three types, *Risky-Choice Framing* involves choices between options that differ in their level of risk, where framing outcomes as gains or losses typically shifts risk preference (Tversky & Kahneman, 1981). *Attribute Framing*, in contrast, focuses on how the evaluation of a single object is altered by labeling its characteristics with positive or negative terms (Levin & Gaeth, 1988). *Goal Framing* centers on the persuasive power of a message to encourage a behavior by emphasizing either the positive consequences of acting or the negative consequences of not acting. Given that the current study involves both the evaluation of specific system *attributes* (its performance) and the overarching *goal* of selecting the best system for a task, attribute and goal framing are important theoretical lenses for our investigation. Specifically, attribute framing effects may occur when a single attribute or characteristic of an object or event is framed positively or negatively (Levin et al., 1998). An example is the presentation of meat as either 75 % lean or 25 % fat. In general, attributes are evaluated more positively when they are tagged with positive terms than when they are tagged with negative terms (e.g., Levin & Gaeth, 1988; Liu et al., 2020).

Goal framing effects may occur when the *consequences* of a behavior are framed positively or negatively (Levin et al., 1998). A positive frame focuses attention on the goal of obtaining positive consequences (or gain), whereas a negative frame focuses attention on avoiding negative consequences (or loss). A typical example is highlighting the positive consequences of health screening (e.g., increased chance of finding a disease in an early stage) versus the disadvantages of conducting no health screening (e.g., decreased chance of finding a disease in an early stage). An important aspect of goal framing is that both framing conditions try to promote the same action (e.g., engage in health screening). A negative framing of a message tends to more strongly influence certain behavior (Levin et al., 1998). For instance, a negative framing (you may lose your foot when you do not stop smoking) can more strongly influence certain behavior (stop smoking).

Framing AI-based systems. Research examines framing effects in the field of AI-based systems. For instance, framing affects the acceptance of system recommendations when solving a decision-making task (Kim & Song, 2020). In the respective study, people had to solve a problem together with an AI-based system and system was described as having an 80 % accuracy rate (positive) or a 20 % error rate (negative). Participants were more likely to accept the outputs of the AI-based system (i.e., changed their opinion more in line with the AI-based

[2] Research questions and hypotheses were preregistered under aspredicted.org/KY2_R23.

system's suggestion) when the information was framed positively than when it was framed negatively. However, this study showed even higher acceptance in a condition receiving no information about system accuracy compared to both framing conditions. In another study, the recommendations of a decision aid were more likely to be accepted when the reliability of the decision aid was framed positively (correct in 8 out of 10 cases) than when the description followed a negative framing (incorrect in 2 out of 10 cases; Septiari & Goedono, 2019).

In the case of the accuracy-transparency trade-off, we decided to frame the performance attribute of the systems either positively (describing it as system accuracy) or negatively (describing it as system error rate). First, this could lead to a goal framing effect because the framing of the performance attribute may influence the extent to which decision-makers' value system accuracy within the accuracy-transparency trade-off. Specifically, framing system performance positively (negatively) could make it comparatively more salient that the goal of the system is to maximize accuracy (minimize error). As a result, a comparatively stronger salience of one of the respective goals could influence people's choice between systems whose properties may be perceived as differently successful in achieving a gain versus preventing a loss (Levin et al., 1998).

We propose that there may be a shift towards preferring high accuracy over transparency when system performance is presented in a negative frame. Describing system performance as error rate may lead to people focusing on possible system errors. This could bring decision-makers into a loss-averse way of decision-making (Kahneman & Tversky, 1979), which should reduce the willingness to accept a higher likelihood of errors in favor of more transparency.

Hypothesis 2. The presentation of the system's performance in a negative frame (i.e., as error rate) is associated with a system choice with a preference for accuracy.

Second, framing system performance as accuracy versus error rate may lead to an attribute framing effect as the manipulation of the performance attribute can affect the perception of the systems. Specifically, framing the performance attribute as error rate versus accuracy rate may diminish the perceived competence of the AI. People expect automated systems to work near-to-perfection and violating this expectation of perfection can diminish people's perceived trustworthiness and trust in system outputs (Dietvorst et al., 2015; Madhavan & Wiegmann, 2007). Describing the performance of a system with its error rate may violate expectations of perfection as the term "error rate" makes it salient that the system has already produced errors in the past and will produce errors in the future. This may reduce the perceived competence of the system.

Perceived trustworthiness in general and perceived competence of a system as a facet of trustworthiness are crucial for trust development in general (Baer & Colquitt, 2018). Given that framing as error rate reduces perceived competence of systems, this may ultimately also reduce perceived trust in the system.

Hypothesis 3. The a) perceived competence of the system and b) trust is higher when the system's performance is presented in a positive frame (accuracy) than in a negative frame (error rate).

2.2.3. Accountability

Accountability, in its broadest psychological sense, refers to the implicit or explicit expectation of being asked to justify one's beliefs, feelings, or actions to others (Lerner & Tetlock, 1999, p. 255). Accountability often affects decision-making, in particular by improving information gathering and processing as well as by stimulating a more thorough decision-making process (Aleksovska et al., 2019). In our context, there is no definitive right or wrong choice, as we deliberately clarified that all available systems are market-ready. This approach allows us to focus on understanding preferences in system choice, ensuring that a more deliberate decision-making process does not inherently influence which system is ultimately chosen. Nevertheless, in line with earlier research, accountability may lead to less risky decision-making, which could go hand in hand with an error-minimizing/accuracy-optimizing strategy (Polman & Wu, 2020).

Research Question 1. Does accountability for their decision influence participants' system choice in terms of preference for transparency or performance?[3]

2.2.4. Context of use

The context of use may affect preferences for transparency versus accuracy (e.g., Woodruff et al., 2020). For example, Nussberger et al. (2022) showed that people prioritize accuracy over interpretability in contexts where the stakes are high. Furthermore, initial qualitative research implies that in medical contexts, people may prioritize accuracy at the expense of possible transparency, whereas in contexts related to hiring or criminal justice, transparency seems to be considered as more important even at the expense of accuracy (van der Veer et al., 2021). Further qualitative research has produced similar findings, where respondents reported that they are less concerned about the transparency of a system in medical use cases compared to hiring contexts (Singh, 2019). Furthermore, individuals might expect that no matter how transparent decision processes are, in the medical field they do not have the expertise to understand the decision process. People may further distinguish between contexts where explanations are expected from human decision-makers (e.g., hiring or credit scoring) and contexts where explanations are less expected (e.g., medical contexts; van der Veer et al., 2021). Such expectations could also be applied in the context of AI-based systems and may explain varying preferences regarding accuracy and transparency between different contexts of use. Taken together, these qualitative findings suggest that laypeople's preferences for performance versus transparency vary systematically across domains, providing a rationale for the quantitative investigation undertaken in this study.

Research Question 2. Does the context of use influence the preference for performance or transparency when choosing a system, with participants preferring performance in a medical context of use?

2.2.5. Exploratory: effects of person-related factors on system choice

Choosing between different available AI-based systems may be affected by person-related factors and individual differences. We chose to examine two person-related factors that may be central to individual preferences for accuracy/transparency: propensity to trust and affinity for technology interaction (ATI). These two factors represent a general disposition toward technology (propensity to trust) and an active engagement style when confronted with technology (ATI). A person's propensity to trust reflects the extent to which this person generally trusts systems (Merritt & Ilgen, 2008). In terms of the influence on the system choice, a high propensity to trust comes with a higher willingness to rely on a system and its outputs. This higher willingness will be comparably easier for systems that have a low probability for errors. Thus, a high propensity to trust may be associated with a preference for

[3] In the preregistration, we hypothesized a relationship between accountability and a preference for system transparency. However, the theoretical basis of this assumed relationship was that participants would be inclined to prefer a transparent system when they have to explain future decisions of the system. In the study, participants took the role of a deployer and were instructed to choose one of the systems and were informed that they may be asked to justify their choice for one of the systems. Thus, they may not have imagined themselves working together with the system in the future. In hindsight, the actual study context thus differed from our initial theoretical idea, so we decided to alter the specific hypothesis to a research question. In the preregistration, there were also further hypotheses for accountability that were based on the original idea. We decided not to include them in the manuscript. For none of these hypotheses we found significant results.

systems with the highest possible performance. In contrast, a low propensity to trust is associated with a person being more skeptical about systems. A low propensity to trust also comes with the desire to be able to control and monitor systems (Lee & See, 2004), which might be easier for systems with a higher level of transparency.

Affinity for technology interaction (ATI) is defined as "the tendency to actively engage in intensive technology interaction, as a key personal resource for coping with technology" (Franke et al., 2019, p. 1). ATI is closely related to people's need for cognition. Individuals with a strong need for cognition enjoy and actively engage in cognitively demanding activities. Individuals with a strong level of ATI may want to engage with systems and its decisions, might want to understand how systems come to their decisions, reflecting a desire for deep cognitive engagement, and may thus favor transparent systems. However, given the lack of prior research to build hypotheses upon, we decided to pose the following exploratory research question:

Exploratory Research Question 1. How do propensity to trust and affinity for technology interaction relate to the choice between the systems?

3. Method

3.1. Sample

The sample size for this study was determined based on rules of thumb by Long (1997) and Taylor et al. (2006) for ordinal logistic regressions. For the main study, we planned to collect data from a total of about 500 participants.

We collected data via the participant recruitment platform Prolific. To ensure data quality, participation was restricted using Prolific's prescreening tools to those with a 100 % approval rate and a minimum of 10 previous submissions. Completion of the study took an average of 7.27 min and participants received 1.20 British pounds. We collected data from 539 participants. Of these, 20 were excluded because they indicated that their answers should not be used for analysis, 50 because they failed the attention checks regarding the framing or context of use condition (see Attention Check Section),[4] and one person because they filled out the questionnaire in less than 2 min indicating inattentiveness. The final sample consisted of 468 participants, of whom 317 identified themselves as female (67.74 %), 145 as male (30.98 %), and 6 as diverse (1.28 %). The mean age was 36.08 years (range: 18–80; $SD = 12.76$). Most participants were from the United Kingdom (64.94 %), with the remaining participants from the United States (8.44 %), South Africa (8.23 %), Canada (7.14 %), Ireland (3.46 %), and a small percentage of participants from other countries. Of those participants, 68.38 % were employed and 16.67 % stated that they were currently studying. The highest level of education reported by 4.73 % of participants was that they attended high school, 18.06 % reported a high school diploma, 18.06 % reported a degree after 2-year community/technical/professional/trade college, 42.80 % reported a 4-year college or university degree, and 16.34 % reported a graduate degree/PhD. All data, analysis code, codebook, and materials are available at https://osf.io/cbd8z/?view_only=3e0341d99920437986d21095761b4672.

[4] We decided to exclude participants who failed the attention checks regarding framing and context of use, since they were hypotheses-relevant and rather obvious manipulations. Failing these attention checks indicate inattentiveness. In contrast, the accountability manipulation was subtle. In line with this, in total 106 participants failed this attention check (of which 21 also failed one of the other attention checks). Given that the results for the accountability manipulation changed (from significant to not significant) when excluding participants who failed the accountability attention check we judged any accountability effects to not be robust. We decided to report analyses for the total number of participants (without filtering for the accountability attention check) and to not interpret the results for the manipulation for accountability.

3.2. Task and contexts

Participants were put in the perspective of a deployer, tasking them with choosing between different AI-based systems for a specific context of use. We selected four different contexts of use of AI-based systems: medicine, hiring, credit scoring, and criminal justice. All scenarios represent high-risk contexts to ensure that trading off between performance and transparency is perceived as important. Similar to the description of available systems the vignettes were guided by available materials from prior studies conducted in the context of AI-based systems (Longoni et al., 2019; Oswald, 2019; van der Veer et al., 2021; Woodruff et al., 2020).

As a medical context, we selected the use of an AI-based system that diagnoses the risk of developing age-related macular degeneration (AMD), a disease of the retina of the eye, based on the images of an ocular imaging procedure. Since ophthalmology is a field that relies on medical imaging for analysis and diagnosis, and since the use of AI-based image classification in ophthalmology seems promising, the use of an AI-based system to diagnose AMD can be seen as a representative context of use from the medical field (Perepelkina & Fulton, 2021). In medicine it is crucial for AI-based systems to function accurately and at the same time to ensure transparency, highlighting the importance of the accuracy-transparency trade-off (e.g., Heinrichs & Eickhoff, 2020).

In the area of hiring, companies increasingly adapt AI-based systems to identify qualified applicants (Gonzalez et al., 2019; Hickman et al., 2025). Therefore, for this context of use, a situation was described where an AI-based system is used to screen applicants. In hiring, accuracy is important as organizations want to select the best applicants for a given position. At the same time, transparency is important because hiring managers as well as applicants may want to know reasons for the outputs of AI-based systems (e.g., why an applicant has been rejected).

In the credit scoring context, the vignette described the use of an AI-based system to optimize the prediction of repayment probability. Credit scoring is a field where the use of statistical models has always been important and, at the same time, the statistical methods from which to choose are increasingly influenced by the accuracy-transparency trade-off (Demajo et al., 2021). Similar to the hiring scenario, loan givers and loan applicants want the most accurate outputs from systems and at the same time may want to understand why an applicant was approved/rejected.

Lastly, the criminal justice vignette portrayed an AI-based system that would be used to select offenders who would be offered a place in a rehabilitation program. The use of algorithmic decision-support in criminal justice is increasing, there are similar tensions between accuracy and transparency when employing algorithmic systems as for the other context of use in this study (Rudin & Ustun, 2018). The full vignettes for each context can be found in Appendix A.

3.3. Procedure

The study was conducted via the online questionnaire platform SoSci Survey and followed a 4 (context of use: medicine vs. hiring vs. credit scoring vs. criminal justice) x 2 (positive vs. negative framing) x 2 (accountability vs. no accountability) between-person design. Fig. 1 shows a flow chart summarizing the study procedure. At the beginning of the study, participants were randomly assigned to one of the sixteen conditions. Participants were first informed about the specific context of use for which they would be choosing an AI-based system.

To experimentally manipulate accountability, we varied the task description in the vignette. In the accountability condition, participants were instructed to make the decision conscientiously, keeping in mind that they alone are responsible for it and must subsequently be able to justify that decision. Participants in the non-accountability condition were also informed that they should make their decision conscientiously, but their choice would only factor into the final decision, and they would not have to justify it.

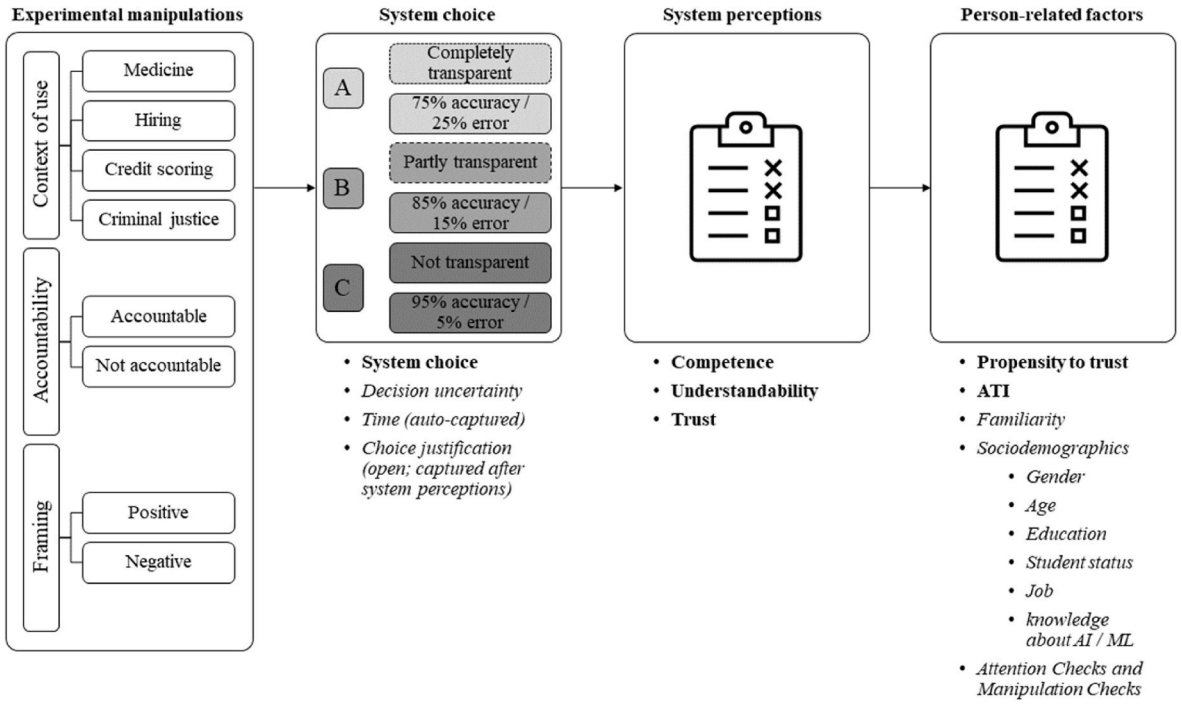

Fig. 1. Flow Chart of the Study Procedure
Notes. Items below the boxes indicate the measured variables. Hypotheses-relevant measures are bolded, exploratory measures are italicized.

Participants were then instructed to choose one of the available systems (with the neutral names A, B, and C). The use of neutral labels was a deliberate choice to avoid potential confounding effects that could arise from pre-existing associations with technology or brand names. To inform this decision, participants received descriptions of three different systems. The description of these systems realized the trade-off between accuracy and transparency. Fig. 2 shows a sample of these descriptions.

We utilized an abstract form of system properties to make the trade-off between accuracy or transparency salient. To ensure that all systems would be perceived as performing well enough to be considered a viable alternative, the vignettes described that different systems are available on the market and that all are already in use in similar use cases.

We framed the performance attribute of the eligible systems either positively or negatively. In the positive framing, the system performance was presented as 75 %, 85 %, and 95 % accuracy rate, respectively. In the negative framing, the system performance was as 25 %, 15 %, and 5 % error rates, respectively. Because statistical information is more difficult to process when presented in probabilistic form, frequencies rather than probabilities should be used to communicate such information (Griffin & Buehler, 1999). Therefore, in addition to the respective percentages, we described these values as relative frequencies (e.g., "It has an overall accuracy rate of 85 %. This means that in 100 cases, a correct prediction is made 85 times on average."). The vignettes and experimental manipulations were also designed to be as parallel as possible in terms of word count and information density to minimize potential differences in decision time attributable to reading duration.

The vignettes including the experimental manipulations and the descriptions of the eligible systems appeared together on one page. Participants confirmed the choice of a system by selecting it and by proceeding to the next page of the questionnaire. On the next page, participants responded to items capturing their perceived trustworthiness and trust in the chosen system, as well as exploratory questions about familiarity with similar systems. Then we captured participants' propensity to trust in AI-based systems and their ATI. Across the questionnaire, we used several attention and manipulation checks. In the end, we asked participants to provide demographic information (gender, age, education level, student status, job) and their knowledge about AI or machine learning.

3.4. Measures

Unless otherwise stated, participants responded to all items measuring the dependent variable on a scale from 1 (*strongly disagree*) to 5 (*strongly agree*).

3.4.1. System choice

The main dependent variable was the participant's system choice. This was a behavioral preference measure where participants selected one of three systems (System A, B, or C). The systems were designed to operationalize the accuracy-transparency trade-off, with System A offering the highest transparency and lowest accuracy, and System C offering the lowest transparency and highest accuracy. For the purpose of ordinal regression analysis, this categorical choice was coded as an ordered variable (1 = System A, 2 = System B, 3 = System C), where higher values represent a stronger preference for accuracy over transparency.

3.4.2. Perceived trustworthiness and trust

For measuring perceived trustworthiness with its facets competence and understandability, as well as perceived trust in the system, and propensity to trust we used validated scales from Körber (2019). Competence was captured with six items (sample item: "I am confident in the capabilities of the system."). Understandability was captured with four items which were adapted to the context of the present study to reflect that participants were making a judgment before any direct interaction with the system (sample item: "I could understand why something happened."). Trust in the system was measured with two items (sample item: "I trust the system."), and propensity to trust with three items (sample item: "I trust a system rather than distrust it.").

3.4.3. Affinity for technology interaction (ATI)

Affinity for technology interaction (ATI; Franke et al., 2019) was measured with the short version (ATI-S) of the validated scale. The ATI scale is theoretically grounded in the *Need for Cognition* construct and

System A:

This system is **completely transparent** in the way it reaches its conclusions: for each individual case it can provide specific rules that were applied to reach a conclusion. The explanations the system provides are fully understandable by humans.

It has an overall **accuracy rate of 75%**. This means that in 100 cases, a correct prediction is made 75 times on average.

System B:

This system is **partly transparent** in the way it reaches its conclusions: it can only tell us which features, i.e., which variables or attributes, in general, are important and which are not. Thus, the explanations the system provides are only partly understandable by humans.

It has an overall **accuracy rate of 85%**. This means that in 100 cases, a correct prediction is made 85 times on average.

System C:

This system is **not transparent** in the way it reaches its conclusions: it is unable to provide any explanation that is understandable by humans.

It has an overall **accuracy rate of 95%**. This means that in 100 cases, a correct prediction is made 95 times on average.

Fig. 2. Description of the available systems (Positive framing condition).

assesses an individual's tendency to actively seek out and engage with technology on a deeper level. The short scale uses four of the nine items in the original scale (sample item: "I like to occupy myself in greater detail with technical systems.") The items were assessed using the original scale ranging from 1 (*not at all true*) to 6 (*completely true*).[5]

3.5. Attention checks and manipulation check

To ensure data quality, a general attention check question was included in the questionnaire (Oppenheimer et al., 2009): "Check the response option 'Strongly agree'." We also included attention checks concerning the experimental manipulations. These condition-specific checks were deliberately placed at the end of the survey to avoid influencing participants' responses during the main decision task. First, we asked "For which application area was the AI system you selected intended to serve?" and presented the four contexts of use as possible answers (medicine, hiring, credit scoring, or criminal justice). The attention check for framing was "In which way was the prediction performance of the systems presented to you?" (accuracy rate or error rate). Finally, we asked participants if they were asked to feel accountable within this study.

To test whether the manipulation within the vignette induced more perceived accountability in the accountability condition than in the non-accountability condition, we used two items from Zhang and Mittal (2005), originally developed by Lerner et al. (1998). The items were answered on a five-point scale (sample item: "How concerned were you regarding a possible bad decision?", 1 (*not at all concerned*) to 5 (*very concerned*)).

[5] For exploratory purposes, after participants decided for a system, we also assessed decision uncertainty with three items from the Decisional Conflict Scale (DCS) by O'Connor (1995). We asked participants to indicate their decision uncertainty regarding their choice and gave them the opportunity to provide reasons for their decision for the selected system for the corresponding context of use in an open response format. Additionally, participants could state how important the performance and transparency of a system is to them in the given situation, irrespective of the available systems. This was done to gain an impression of the extent to which the accuracy-transparency trade-off is reflected in participants' preference. We also captured familiarity with systems with two items by Körber (2019).

4. Results

Fig. 3 shows the descriptive statistics of system choice in terms of absolute as well as relative frequencies for the 16 groups of the 2 x 2 x 4 design. Table 1 shows the descriptive statistics and reliabilities of the study variables. Also, Table 1 shows the correlations between all variables as well as with the dichotomous experimental conditions framing and accountability.

4.1. Research questions and hypotheses

The dependent variable, system choice, was ordinally coded for this purpose, in the sense that higher values indicate a preference for a system with greater accuracy/lower transparency, and lower values indicate a preference for a system with lower accuracy/higher transparency. In the regression results, a significant positive *b* weight means that the corresponding predictor variable increases the likelihood of choosing accuracy over transparency compared to the reference category. Conversely, a negative *b* weight suggests a preference for transparency over accuracy.

4.1.1. Effects on system choice

The common analysis for testing Hypothesis 2, Research Question 1, and Research Question 2 is an ordinal logistic regression model that includes the experimental groups as predictors (framing, accountability, and context of use) and the ordinal dependent variable system choice. Table 2 shows the resulting ordinal logistic regression model. Overall, the regression model was significant, likelihood-ratio-$\chi^2(5) = 15.01$, $p = .010$. The Nagelkerke's R^2 of 0.036, which serves as the model-level indicator of effect size, suggests the model explained approximately 4 % of the variance in system choice. As shown in Table 2, accountability was a significant predictor of system choice, likelihood-ratio-$\chi^2(1) = 4.69$, $p = .030$. The resulting Odds Ratio ($OR = 1.45$, 95 % CI [1.04, 2.03]), the indicator of effect size for this type of analysis, showed that participants in the accountable condition were 1.45 times more likely to prefer accuracy. The context of use also significantly affected system choice, likelihood-ratio-$\chi^2(3) = 9.20$, $p = .027$. Using medicine as the reference category, the odds of preferring transparency were significantly higher in the hiring context ($OR = 0.53$, 95 % CI [0.33, 0.86]), the credit scoring context ($OR = 0.53$, 95 % CI [0.33, 0.86]), and the criminal justice context ($OR = 0.60$, 95 % CI [0.37, 0.96]). In other words, the probability of choosing a system with a preference for performance was approximately 1.88 times higher in the medical context

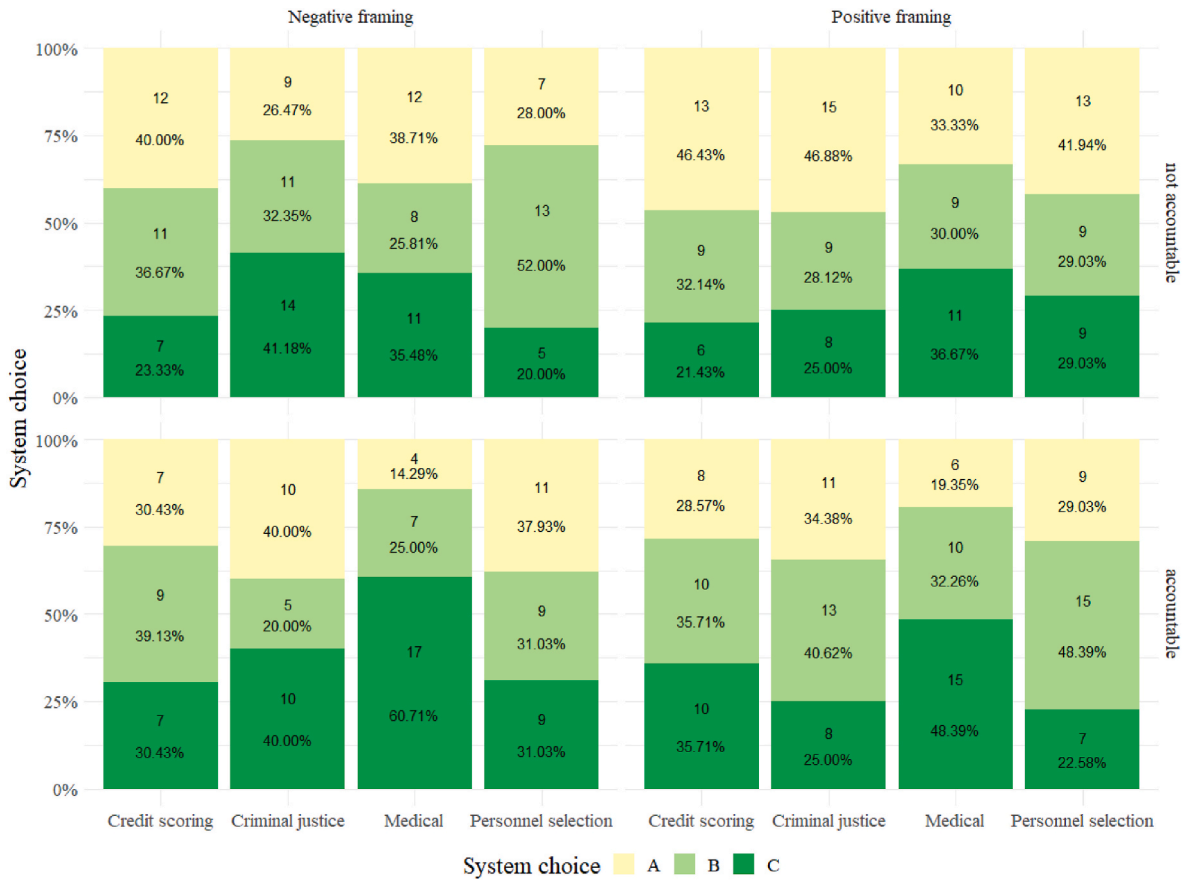

Fig. 3. Distribution of system choice by framing, accountability, and context.
Notes. Absolute and relative frequencies of the system choice. System A corresponds to 75 % accuracy/25 % error and full transparency. System B corresponds to 85 % accuracy/15 % error and partial transparency. System C corresponds to 95 % accuracy/5 % error and no transparency. $N = 468$.

Table 1
Descriptive statistics, reliabilities and correlations of the study variables.

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11
Conditions													
1. Framing	–	–											
2. Accountability	–	–	.06[a]										
System choice													
3. System choice	1.99	0.82	−.07[a]	.14[a]**									
4. Time (System choice)	131.30	123.79	−.06[b]	−.04[b]	.00[b]								
5. Decision uncertainty	3.15	1.02	−.05[c]	.06[c]	−.00[d]	.07[b]	(0.85)						
System perception													
6. Competence	3.49	0.60	.20[c]*	.05[c]	.53[d]**	.09[b]*	−0.12**	(0.77)					
7. Understandability	3.33	0.87	.07[c]	−.04[c]	−.68[d]**	−.02[b]	−0.16**	−0.09	(0.75)				
8. Trust	3.72	0.77	.15[c]*	.03[c]	.27[d]**	.11[b]*	−0.20**	0.68**	0.10*	(0.87)			
Person-related factors													
9. Propensity to trust	2.80	0.59	.07[c]	.03[c]	.33[d]**	−.03[b]	−0.08	0.45**	−0.10*	0.48**	(0.55)		
10. Familiarity	1.90	0.93	.05[c]	.03[c]	−.00[d]	−.01[b]	−0.09	0.08	0.12**	0.14**	0.16**	(0.90)	
11. ATI	3.49	1.01	.02[c]	−.07[c]	−.17[d]**	.09[b]*	−0.10*	−0.01	0.14**	0.06	−0.01	0.25**	(0.79)
12. AI-Knowledge	2.34	0.93	.08[c]	.02[c]	−.08[d]	.11[b]*	−0.14**	0.09	0.14**	0.07	0.11*	0.35**	0.54**

Notes. AI = Artificial intelligence, ATI = Affinity for technology interaction. Unless otherwise indicated, Pearson product-moment correlations are shown. Time variable in seconds. Coding of framing: 0 = negative, 1 = positive. Coding of accountability: 0 = not accountable, 1 = accountable. Coding of system choice: lower value = preference for transparency, higher value = preference for accuracy. The diagonal shows in brackets the internal consistencies (*Cronbach's* α). $N = 468$. *$p < .05$. **$p < .01$.
[a] Polychoric correlations.
[b] Spearman rank correlations.
[c] Pearson point-biserial correlations.
[d] Polyserial correlations.

Table 2
Results of the ordinal logistic regression with system choice as criterion.

Variable/Category		*b*	*SD*	*t*	*p*	*OR* (95 % CI)	χ^2	*df*	*p*
Framing[a]							1.46	1	0.226
Positive		−0.21	0.17	−1.21	0.227	0.81 [0.58, 1.14]			
Accountability[b]							4.69	1	0.030*
Accountable		0.37	0.17	−2.16	0.031*	1.45 [1.04, 2.03]			
Context of use[c]							9.20	3	0.027*
Credit scoring		−0.63	0.24	−2.57	0.010*	0.53 [0.33, 0.86]			
Criminal justice		−0.51	0.24	−2.13	0.033*	0.60 [0.37, 0.96]			
Hiring		−0.63	0.24	−2.60	0.009**	0.53 [0.33, 0.86]			
Intercept A|B		−1.07	0.22	−4.91	<0.001**				
Intercept B|C		0.36	0.21	1.68	0.924				
Residual deviance	1013.19								
AIC	1027.19								
LogLik	−506.60								

Notes. b = unstandardized regression weights. The Odds Ratio (*OR*) column, which includes the 95 % Confidence Interval (CI), serves as the primary indicator of effect size for each predictor. AIC = Akaike information criterion. LogLik = Log-Likelihood. $N = 468$. *$p < .05$. **$p < .01$.

[a] Reference category = negative.
[b] Reference category = not accountable.
[c] Reference category = medicine.

compared to the hiring or credit scoring contexts. Framing ($p = .226$) was not a significant predictor.

Hypothesis 2 proposed that presenting the performance of the systems in a negative frame as an error rate would be associated with a system choice with a preference for performance. In the regression model, the main effect for framing was not significant, likelihood ratio-$\chi^2(1) = 1.46$, $p = .226$. The Odds Ratio, our indicator of effect size, showed that the positive frame had a negligible effect on system choice compared to the negative frame ($OR = 0.81$, 95 % CI [0.58, 1.14]). Thus, Hypothesis 2 was not supported.

Research Question 1 asked whether participants are more likely to show a preference for transparency if they have to be accountable for their decision. In the regression model, the main effect accountability was significant, likelihood ratio-$\chi^2(1) = 4.69$, $p = .030$. The Odds Ratio ($OR = 1.45$, 95 % CI [1.04, 2.03]), an indicator of effect size, initially suggested that participants in the accountable condition were more likely to prefer accuracy. Consequently, the response to Research Question 1 is that accountability might affect system choice. However, in robustness analyses where we excluded participants who failed the attention check for accountability, the effect was unstable (i.e., the significance depended on which participants we excluded or included, see also Footnote [3]). We therefore decided not to interpret the effect on accountability.

Research Question 2 suggested that the context of use influences the preference for performance or transparency when choosing a system, with a preference for performance in a medical context. In the regression model, the main effect for the context of use was significant, likelihood ratio-$\chi^2(3) = 9.20$, $p = .027$. To allow for the comparison of the medical context with all other contexts, we used the medical context as the reference category for the dummy variables of the categorical predictor context of use. Table 2 shows that compared to the medical context, there was a stronger tendency to choose systems offering more transparency in ($OR = 0.53$, 95 % CI [0.33, 0.86]), criminal justice ($OR = 0.60$, 95 % CI [0.37, 0.96]), and hiring contexts ($OR = 0.53$, 95 % CI [0.33, 0.86]). The probability of making a system choice with a preference for performance was approximately 1.88 times higher (1/0.531) in a medical context compared to a credit scoring context, approximately 1.67 times higher (1/0.598) compared to a criminal justice context, and approximately 1.88 times higher (1/0.533) compared to a hiring context. The response to Research Question 2 is that the context of use influences the preference for performance or transparency when choosing a system, with participants preferring performance in a medical context compared to other contexts of use.

4.1.2. Exploratory: effects of person-related factors on system choice

Exploratory Research Question 1 asked whether propensity to trust and affinity for technology relate to system choice. We computed an ordinal logistic regression model where we included all investigated variables, propensity to trust and ATI, and interactions between experimental variables. In the final regression model, the main effect of the context of use was still significant, likelihood ratio-$\chi^2(3) = 10.17$, $p = .017$. In addition, affinity for technology ($p < .001$), and propensity to trust ($p < .001$) significantly affected system choice.

Higher propensity to trust was related to a higher likelihood of a system choice with a preference for performance, $b = 1.07$, $SD = 0.16$, $OR = 2.15$, 95 % *CI* [2.12, 4.03], $p < .001$. Higher ATI was associated with a higher likelihood of system choice with a preference for transparency, $b = -0.32$, $SD = 0.09$, $OR = 0.73$, 95 % *CI* [0.61, 0.87], $p < .001$.

4.1.3. Effects on system perceptions

The effects on system perceptions were tested using several linear regression models.[6] Hypothesis 1a proposed that a system choice with a preference for transparency would be associated with higher levels of perceived understandability. The respective regression was statistically significant, $R^2 = 0.39$, $F(2, 465) = 150.12$, $p < .001$. System choice was significantly associated with perceived understandability: the difference between System A and B, $b = -0.75$, $p < .001$, as well as between B and C, $b = -0.76$, $p < .001$, were significant. The negative regression weights imply that a system choice with a tendency for transparency was associated with stronger perceptions of understandability. This supports Hypothesis 1a.

Hypothesis 1b proposed that a system choice with a preference for performance will be associated with higher levels of perceived competence. The respective regression was statistically significant, $R^2 = 0.23$, $F(2, 465) = 67.77$, $p < .001$. Perceived competence was predicted by

[6] In order keep to the ordinal structure of the variable system choice when used as a predictor, a special coding scheme for ordinal independent variables in multiple regressions was used for the linear regression models (Walter et al., 1987). The codes are chosen in such a way that, in the case of three levels, the first dummy variable encodes the difference between the first and the second level and the second dummy variable encodes the difference between the second and the third level. Thus, the dummy variables can be interpreted not only as a function of a reference category, as in classical dummy coding, but also represent the ordinal structure of the variables, since two pairwise comparisons are possible. Positive *b*-weights of the dummy variables indicate a tendency for accuracy and negative *b*-weights of the dummy variables indicate a tendency for transparency.

system choice: the difference between System A and B, $b = 0.17$, $p = .006$, as well as between B and C, $b = 0.51$, $p < .001$, were significant. The positive regression weights imply that a system choice with a tendency for performance was associated with increased perceived competence. Hypothesis 1b was thus supported.

Hypothesis 3a suggested that the perceived competence of the system is higher when the performance of the system is represented in a positive frame as accuracy than when it is represented in a negative frame as error rate. To test whether framing significantly predicts perceived competence when controlling for system choice (we controlled for system choice because we found a relation between system choice and perceived competence in Hypothesis 1b.), we used a multiple linear regression model. The regression was statistically significant, $R^2 = 0.26$, $F(3, 464) = 53.95$, $p < .001$. Framing significantly predicted perceived competence, $b = 0.17$, $p < .001$ in a way that positive framing was associated with a higher perceived competence compared to negative framing. Hypothesis 3a was thus supported.

Hypothesis 3b proposed that trust in the system is higher when the system's performance is framed as accuracy in a positive frame than when it is framed as error rate in a negative frame. A linear regression was used to test whether framing significantly predicted trust. The regression was statistically significant, $R^2 = 0.01$, $F(1, 466) = 6.39$, $p = .012$. Framing significantly predicted trust, $b = 0.18$, $p = .012$ in a way that positive framing was associated with higher trust compared to negative framing. Hypothesis 3b was thus supported.

5. Discussion

This study examined how individuals navigate the choice between AI-based systems within the accuracy-transparency trade-off. In an experimental design, participants were in the role of a system deployer and selected from systems whose properties reflected this trade-off. We manipulated three key factors: the framing of performance, the accountability of the decision-maker, and the context of use, analyzing their impact on system choice. Additionally, we assessed whether system choice and the framing of system properties influenced perceptions of the systems. Lastly, we explored the role of person-related factors in shaping system preferences.

The results for system choice revealed a stronger preference for performance in medical contexts compared to other settings. While accountability significantly influenced system choice, these effects were not robust therefore not interpreted further. Framing had no significant impact on system choice. Additionally, person-related factors played a role: a higher propensity to trust was linked to a preference for performance, while a higher affinity for technology was associated with a preference for transparency. As anticipated, choosing a system favoring transparency was associated with higher perceived understandability, whereas a preference for performance corresponded to higher perceived competence. Lastly, framing an AI system's performance as accuracy rather than error rate enhanced perceptions of the system's competence and trust in the system.

5.1. Theoretical implications

5.1.1. Effects of context of use, framing, and accountability on system choice

As expected, participants in the role of a deployer showed a significant preference for performance in deciding between different available systems for the medical context. Undoubtedly, transparency of AI-based systems is relevant in the medical domain: to enable justification and contestability of diagnoses and treatment decisions, as well as to be able to get new clinical insights (He et al., 2019). However, our results suggest that, in contrast to other contexts of use, our participants preferred accuracy over transparency in the medical context. Those results extend initial findings from qualitative research that indicated that accuracy may be valued more than transparency when using AI-based systems in medicine compared to other high-stakes contexts (e.g., van der Veer et al., 2021).

There are several possible reasons for this finding. First, the preference for accuracy is likely driven by heightened risk perceptions; a higher perceived severity of consequences in medicine may lead to a stronger focus on risk mitigation through maximizing accuracy (Nussberger et al., 2022). Second, transparency – also for human physicians – may be less expected in the medical context. People may expect that even if decisions are made transparent in the medical domain, they may lack the knowledge to understand the information or act upon this information, thereby undermining the potential for meaningful contestability. This suggests a willingness to delegate expertise and consequently accept more opacity when the domain is highly specialized and consequences of errors are potentially dire. In other words, improving transparency without expecting improved understanding and control may not justify losses in system performance (Ananny & Crawford, 2018; Grote, 2023). Third, in medicine, the impact of low transparency may be less obvious or perceived as less important than low performance, especially when choosing between accuracy and transparency.

We did not find evidence for the hypothesized effect of negative framing on system choice. Although the results indicated a trend consistent with the hypothesis (i.e., a preference for performance in the negative framing condition), it was not significant. Thus, the effect may have been smaller than expected, or other influences (situational or on an individual level) mitigated the framing effect (Levin et al., 1998). An alternative interpretation is that the assumed goal framing did not lead to a stronger desire for performance, but only increased the saliency of possible errors (see, e.g., Byrne & Hart, 2009). In other words, in the error framing condition, participants may have been more likely to think that no matter what system they choose, errors will happen. Thus, participants may not be motivated to maximize performance, because they believe errors are part of all systems. As a result, participants may perceive their system choice to have less influence on the performance.

The results showed that under certain conditions, accountability may influence the preference in system choice. Since this effect turned out not to be robust after further analysis, we refrain from further interpretation of this effect. However, we still believe it makes sense for research to examine the effects of accountability on system choice.

5.1.2. Person-related factors influencing system choice

The decisions of individuals tasked with selecting among different AI-based systems can be influenced by person-related factors and individual differences. In our study, we focused on two key factors: affinity for technology interaction (ATI) and propensity to trust, as they relate to assessing AI system properties within the accuracy-transparency trade-off. We interpret the findings through the lens of a shared underlying need for control (Bennett et al., 2023).

Specifically, exploratory analyses revealed that higher ATI was significantly associated with a greater likelihood of choosing a more transparent system. ATI reflects an interaction style characterized by a desire to understand and engage with technical systems beyond their basic functions (Franke et al., 2019). The concept is theoretically rooted in and closely related to the construct of need for cognition, which describes an individual's propensity for cognitive engagement (Cacioppo & Petty, 1982). Both active interaction with and cognitive engagement in technical systems suggest a desire for control over these systems. Within the context of the accuracy-transparency trade-off, a preference for transparency appears to align with this desire for control over system processes and outcomes. Consequently, individuals with high ATI were more likely to favor transparency over accuracy.

Higher levels of propensity to trust were significantly associated with a greater likelihood of choosing a system with a preference for performance, while lower levels of propensity to trust were linked to a preference for a more transparent system. As with individuals with strong ATI realizing a need for control through higher system transparency, those with low propensity to trust may also express their need for control

by prioritizing transparency. Opting for transparency over performance offers greater opportunities to monitor and oversee the system, such as by understanding its processes and outputs. Conversely, the preference for performance among individuals with high propensity to trust can also be interpreted through the lens of control. In this case, selecting a high-performance system may reflect a form of control rooted in their strategy to trust systems generally. By choosing a system with strong performance, they reduce the likelihood of that trust being undermined, thereby aligning their choice with their confidence in the system's reliability.

In sum, an underlying need for control may explain the preference for accuracy and transparency for different person-related factors (i.e., ATI and propensity to trust) and for different extremes for a single variable (i.e., high and low levels for propensity to trust) – a tentative interpretation that may be worth investigating in future research.[7]

5.1.3. Factors influencing system perceptions

The results supported all hypothesized relationships between system choice and framing with perceived understandability, competence, and trust. A system choice with a preference for performance was associated with perceiving a higher level of perceived competence for the system and a system choice with a preference for transparency was associated with a higher level of perceived understandability of the system. Note that these relationships could also be understood as a manipulation check, as they build the basis for the link between the accuracy-transparency trade-off and trustworthiness facets. However, they also offer valuable insights by linking the preference for one attribute over the other to individuals' perceptions of the system.

The findings also indicate that positively framing a system's performance enhanced perceptions of competence and increased trust compared to framing its performance in terms of error rates. This aligns with prior research on framing effects in AI-based systems (Septiari & Goedono, 2019). There are several explanations for these findings: First, positive or negative framing triggers corresponding associations, which influence subsequent evaluations (Levin et al., 1998). Second, framing directs attention to the highlighted perspective while diverting focus from the opposite frame, leading to evaluations consistent with the emphasized attributes (Kreiner & Gamliel, 2018). In the context of competence and trust, both mechanisms are plausible: focusing attention on accuracy while ignoring possible errors explained why the system would be perceived as more competent. Our findings on attribute framing align with a growing body of research showing that how system information is presented impacts user perceptions and behavior (Spatola, 2024).

5.2. Main practical implications

This study offers insights for a crucial group of stakeholders: the system deployers. Viewing their role through a socio-technical systems lens, deployers act as organizational gatekeepers; their initial choice determines which AI systems are procured and implemented, thereby shaping all subsequent interactions and downstream consequences for end-users and the public. As AI-based systems continue to proliferate, an increasing number of individuals in such gatekeeping roles, including non-experts, will face the challenge of making these high-stakes decisions. Our findings provide empirical guidance for this critical task.

This study aligns with research emphasizing the critical role of the context of use in shaping preferences for accuracy and transparency in AI-based systems (Dong et al., 2024; Nussberger et al., 2022). These findings underscore two key points. First, the context of use significantly influences the prioritization of system properties that often involve a complex trade-off. Second, it directly impacts deployers' choices, as illustrated by the tendency to prioritize accuracy over transparency in medical domains. Such preferences carry important implications: while more accurate systems may reduce error rates, they may also limit understanding of the rare errors that do occur and pose challenges for human oversight of highly accurate yet opaque systems (Green, 2022; Sterz et al., 2024). These results highlight the need to account for the context of use when designing AI systems for environments where stakeholders may have strong preferences for particular attributes. For system designers and providers, this suggests the value of moving beyond a one-size-fits-all approach. Specifically, our findings on person-related factors suggest that even within a single context of use, individuals have different preferences. A key practical implication, therefore, is to design systems that allow for user choice. Rather than deploying a single model optimized for one attribute, providers could offer deployers the ability to select or switch between different models that prioritize different attributes (e.g., one favoring maximal accuracy, another favoring higher transparency). This approach would better accommodate the diverse needs and preferences of the end-users. For policymakers, these findings highlight the need for context-specific guidelines rather than a universal approach to AI regulation. Ultimately, public trust in AI systems could hinge on aligning system properties with the specific demands of their intended context, ensuring the "right" system is chosen along the accuracy-transparency trade-off.

Empirical evidence on the impact of framing system properties on perceived trustworthiness and trust is particularly significant, given the ubiquity of framing in practice. System properties are routinely highlighted in system manuals, technical documentation, websites, or during system audits (Landers & Behrend, 2023). First, system providers could enhance perceived trustworthiness and trust in AI-based systems with minimal effort by framing system performance as accuracy rather than error rate, without altering the system itself. Clearly, this also leads to ethical concerns, given that organizations can influence trust in their systems without changing anything substantial about the system properties. Second, negatively framing system performance could mitigate automation bias and overtrust (Parasuraman & Manzey, 2010) by drawing greater attention to potential errors, thereby encouraging more cautious and critical interactions with the system. However, providers must be careful, as explaining a system's weaknesses can sometimes backfire, leading users to over-correct the system and worsen outcomes even in situations where the system's output is correct (Rieger et al., 2023, 2025).

5.3. Limitations

First, while our sample from Prolific is demographically diverse, it is not composed of professional deployers, which may limit the generalizability of our findings to expert populations. Actual deployers are likely to face greater responsibility for their decisions, possess more information about available systems, and have more time to evaluate options. In larger organizations, deployers may also rely on advisors who provide recommendations about different systems. However, it is worth noting that any individual can become a deployer when using AI-based systems for personal purposes. Moreover, it remains unclear whether current organizational deployers have significantly more experience assessing AI system properties compared to the participants

[7] Readers may have realized that the correlations in Table 1 between the trustworthiness facets competence and understandability as well as trust propensity and trust show partly unexpected patterns. For instance, competence and understandability correlate slightly negative with each other and trust propensity is highly correlated with trust and competence, but significantly negatively correlated with understandability. This unexpected pattern can be explained by examining the correlation patterns separately by system choice (see supplementary tables D1, D2, and D3 in Appendix D). Here, the patterns are more in line with theory. For instance, within each subgroup of participants who chose a respective system, system competence and understandability show positive correlations. Furthermore, there are positive correlations between the propensity to trust and understandability. This indicates that in spite of unexpected correlations for the overall sample, the expected pattern show for each subgroup of the data (cf. Yule-Simpson paradox; e.g., Goltz & Smith, 2010).

in our study. Furthermore, a full exploration of how cultural backgrounds or prior professional experiences might shape these preferences is beyond this paper's scope and represents a key direction for future research. Nevertheless, future research could benefit from focusing on decision-making processes of deployers in larger organizations, especially when tasked with selecting between various AI-based systems in practical, high-stakes contexts.

Second, the sequence of measurements within the experimental procedure may be a further limitation. All variables were recorded after participants had chosen one of the systems. Especially for the trait variables, such as the propensity to trust, familiarity, or affinity for technology, it cannot be ruled out that they were influenced by the system choice. Additionally, with the current design, no causal conclusions can be drawn, for instance, to attribute the system choice to the trait variables. A similar limitation applies to the effects on the perception of system trustworthiness and trust. Since participants reported their perceptions of systems only after they chose a system, it may be possible that participants chose the system because of these perceptions, but it could also be the case that the system choice affected people's perceptions of the system. Nevertheless, we explicitly decided on this order of events (i.e., first choosing a system, then reporting perceptions and person-related factors) because a) we wanted to expose participants to three systems in parallel to make the accuracy-transparency trade-off salient, and b) we did not want to bias our main dependent variable system choice by presenting the questionnaires before the system choice.

A third limitation is the study's reliance on self-report data within hypothetical scenarios. While hypothetical scenarios are a common and necessary method for studying these trade-offs in a controlled manner, they may not fully capture the complexities of real-world decision-making. Future work could aim to complement these findings with behavioral measures, field studies, and longitudinal designs to understand the long-term effects of an initial system choice on user trust and behavior. Finally, the fact that neither framing nor accountability influenced system choice may be related to the fact that there was only a single decision to make. Whether the effects differ in situations where multiple choices must be made remains open to future study. At the same time, the behavior-guiding effect of the context of use, even in the case of a single decision, may underscore the importance of considering the context of use.

5.4. Conclusion

Our study showed that the context of use and the presentation of properties of AI-based systems influence the choice between AI-based systems and the perception of AI-based systems. The results confirmed the preference for performance in medical applications and identified key person-related factors that also influence choices within the accuracy-transparency trade-off. These findings provide several starting points for future research, including deeper investigations into the psychological drivers of context-specific preferences and the practical effects of framing on stakeholder trust. Ultimately, this work underscores that navigating the accuracy-transparency trade-off is not merely a technical problem but a deeply human one, providing an empirical foundation for designing AI-based systems that are better aligned with the values and expectations of the people who use them.

CRediT authorship contribution statement

Tim Hunsicker: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cornelius J. König:** Writing – review & editing, Validation, Resources, Funding acquisition. **Markus Langer:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Ethics approval

The study has been considered a low-risk study and was thus exempt from ethical approval according to the regulation of the first author's ethical review board.

Funding

Work on this article was partially funded by the German Research Foundation DFG in the project 389792660 as part of the Transregional Collaborative Research Center TRR 248 "Foundations of Perspicuous Software Systems" project A6, by the Bundesministerium für Bildung und Forschung (BMBF) in the project "Ophthalmo-AI" (grant 16SV8640), by the VolkswagenStiftung in the project "Explainable Intelligent System" (AZ 98513), and by the Daimler and Benz Foundation as part of the project TITAN – Technologische Intelligenz zur Transformation, Automatisierung und Nutzerorientierung des Justizsystems (grant no. 45-06/24).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Supplementary data to this article (Appendix A-D) can be found online at https://doi.org/10.1016/j.chbah.2025.100216.

Data availability

We made all data, analysis code, codebook, and materials available at https://osf.io/cbd8z/?view_only=3e0341d99920437986d21095761b4672.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160. doi.org/10/gfvb5g.

Aleksovska, M., Schillemans, T., & Grimmelikhuijsen, S. (2019). Lessons from five decades of experimental and behavioral research on accountability: A systematic literature review. *Journal of Behavioral Public Administration, 2*(2), Article 2. https://doi.org/10.30636/jbpa.22.66

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*. , Article 101805. https://doi.org/10.1016/j.inffus.2023.101805

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Assis, A., Véras, D., & Andrade, E. (2023). Explainable Artificial Intelligence—An analysis of the trade-offs between performance and explainability. In *2023 IEEE Latin American conference on computational intelligence (LA-CCI)* (pp. 1–6). https://doi.org/10.1109/LA-CCI58595.2023.10409462

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in Machine Learning for public policy. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 248–266. https://doi.org/10.1145/3531146.3533090.

Baer, M. D., & Colquitt, J. A. (2018). Why do people trust? In R. H. Searle, A.-M. I. Nienaber, & S. B. Sitkin (Eds.), *The Routledge companion to trust* (1st ed., pp. 163–182). Routledge. https://doi.org/10.4324/9781315745572-12.

Bennett, D., Metatla, O., Roudaut, A., & Mekler, E. D. (2023). How does HCI understand human agency and autonomy?. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–18). https://doi.org/10.1145/3544548.3580651

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), Article 2053951715622512. https://doi.org/10.1177/2053951715622512

Byrne, S., & Hart, P. S. (2009). The boomerang effect: A synthesis of findings and a preliminary theoretical framework. *Annals of the International Communication Association, 33*(1), 3–37. https://doi.org/10.1080/23808985.2009.11679083

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Crook, B., Schlüter, M., & Speith, T. (2023). Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In *2023 IEEE 31st international requirements engineering conference workshops (REW)* (pp. 316–324). https://doi.org/10.1109/REW57809.2023.00060

Demajo, L. M., Vella, V., & Dingli, A. (2021). An explanation framework for interpretable credit scoring. *International Journal of Artificial Intelligence and Applications (IJAIA), 12*(1), 19–38. https://doi.org/10.5121/ijaia.2021.12102

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

Dong, M., Conway, J. R., Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2024). Fears about artificial intelligence across 20 countries and six domains of application. *American Psychologist*. https://doi.org/10.1037/amp0001454

EU Agency for fundamental rights. (2019). Data quality and artificial intelligence – Mitigating bias and error to protect fundamental rights. *FRA - European Union Agency for Fundamental Rights, FRA Focus*. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf.

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human-Computer Interaction, 35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Goltz, H. H., & Smith, M. L. (2010). Yule-Simpson's paradox in research. *Practical Assessment, Research and Evaluation, 15*(1), 15. https://doi.org/10.7275/dgcc-jv81

Gonzalez, M., Capman, J., Oswald, F., Theys, E., & Tomczak, D. (2019). "Where's the I-O?" artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions, 5*(3). https://doi.org/10.25035/pad.2019.03.005

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*, Article 105681. doi.org/10/gqpjps.

Griffin, D., & Buehler, R. (1999). Frequency, probability, and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology, 38*(1), 48–78. https://doi.org/10.1006/cogp.1998.0707

Grote, G. (2023). *Shaping the development and use of Artificial Intelligence: How Human Factors and Ergonomics expertise can become more pertinent*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37). https://doi.org/10.1126/scirobotics.aay7120

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: Legal incentives and technical challenges. *Artificial Intelligence and Law, 28*(4), 415–439. https://doi.org/10.1007/s10506-020-09260-6

Hatherley, J., Sparrow, R., & Howard, M. (2024). The virtues of interpretable medical AI. *Cambridge Quarterly of Healthcare Ethics, 33*(3), 323–332. https://doi.org/10.1017/S0963180122000664

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine, 25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping, 41*(6), 1435–1444. https://doi.org/10.1002/hbm.24886

Hickman, L., Langer, M., Saef, R. M., & Tay, L. (2025). Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology, 110*(6), 846–858. https://doi.org/10.1037/apl0001247

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291. https://doi.org/10.2307/1914185

Kim, T., & Song, H. (2020). The effect of message framing and timing on the acceptance of artificial intelligence's suggestion. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems* (pp. 1–8). https://doi.org/10.1145/3334480.3383038

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 13–30). Springer. https://doi.org/10.1007/978-3-319-96074-6_2.

Kreiner, H., & Gamliel, E. (2018). The role of attention in attribute framing. *Journal of Behavioral Decision Making, 31*(3), 392–401. https://doi.org/10.1002/bdm.2067

Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist, 78*(1), 36–49. https://doi.org/10.1037/amp0000972

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1998). Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin, 24*(6), 563–574. https://doi.org/10.1177/0146167298246001

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research, 15*(3), 374–378. https://doi.org/10.1086/209174

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes, 76*(2), 149–188. https://doi.org/10.1006/obhd.1998.2804

Liu, D., Juanchich, M., & Sirota, M. (2020). Focus to an attribute with verbal or numerical quantifiers affects the attribute framing effect. *Acta Psychologica, 208*, Article 103088. https://doi.org/10.1016/j.actpsy.2020.103088

Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology, 23*(3), 253–263. https://doi.org/10.1007/s10676-020-09564-w

Long, J. S. (1997). *Regression models for categorical and limited dependent variables* (Vol. 7). Sage Publications, Inc.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.2307/258792

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Nussberger, A.-M., Luo, L., Celis, L. E., & Crockett, M. J. (2022). Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. *Nature Communications, 13*(1). https://doi.org/10.1038/s41467-022-33417-3. Article 1.

O'Connor, A. M. (1995). Validation of a decisional conflict scale. *Medical Decision Making, 15*(1), 25–30. https://doi.org/10.1177/0272989x9501500105

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Oswald, M. (2019). Artificial intelligence (AI) & explainability Citizens' Juries report. *Citizens' Juries c.i.c.* http://assets.mhs.manchester.ac.uk/gmpstrc/C4-AI-citizens-juries-report.pdf.

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2023). Challenges in deploying Machine Learning: A survey of case studies. *ACM Computing Surveys, 55*(6), 1–29. https://doi.org/10.1145/3533378

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. doi.org/10/fcmx4h.

Perepelkina, T., & Fulton, A. B. (2021). Artificial intelligence (AI) applications for age-related macular degeneration (AMD) and other retinal dystrophies. *Seminars in Ophthalmology, 36*(4), 304–309. https://doi.org/10.1080/08820538.2021.1896756

Polman, E., & Wu, K. (2020). Decision making for others involving risk: A review and meta-analysis. *Journal of Economic Psychology, 77*, Article 102184. https://doi.org/10.1016/j.joep.2019.06.007

Rieger, T., Manzey, D., Meussling, B., Onnasch, L., & Roesler, E. (2023). Be careful what you explain: Benefits and costs of explainable AI in a simulated medical task. *Computers in Human Behavior: Artificial Humans, 1*(2), Article 100021. https://doi.org/10.1016/j.chbah.2023.100021

Rieger, T., Onnasch, L., Roesler, E., & Manzey, D. (2025). Why highly reliable decision support systems often lead to suboptimal performance and what we can do about it. *IEEE Transactions on Human-Machine Systems*, 1–10. https://doi.org/10.1109/THMS.2025.3584662

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems, 33*(6), 673–705. https://doi.org/10.1007/s10458-019-09408-y

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5). https://doi.org/10.1038/s42256-019-0048-x. Article 5.

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *INFORMS Journal on Applied Analytics, 48*(5), 449–466. https://doi.org/10.1287/inte.2018.0957

Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. *Mensch Und Computer, 2021*, 325–329. doi.org/10/gmvkj7.

Schoenherr, J. R., Abbas, R., Michael, K., Rivas, P., & Anderson, T. D. (2023). Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness. *IEEE Transactions on Technology and Society, 4*(1), 9–23. https://doi.org/10.1109/TTS.2023.3257627. IEEE Transactions on Technology and Society.

Septiari, D., & Goedono, G. (2019). Does attribute framing exist in audit decision aid? *Journal of Accounting and Investment, 20*(1), Article 1. https://doi.org/10.18196/jai.2001107

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, Article 102551. doi.org/10/ghqvr4.

Singh, A. (2019). *Democratising decisions about technology: A toolkit*. RSA. https://www.thersa.org/reports/democratising-decisions-technology-toolkit.

Spatola, N. (2024). The efficiency-accountability tradeoff in AI integration: Effects on human performance and over-reliance. *Computers in Human Behavior: Artificial Humans, 2*(2), Article 100099. https://doi.org/10.1016/j.chbah.2024.100099

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *ACM conference on fairness, accountability, and transparency (ACM FAccT '24)* (p. 13). https://doi.org/10.1145/3630106.3659051

Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement, 66*(2), 228–239. https://doi.org/10.1177/0013164405278580

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458. https://doi.org/10.1126/science.7455683

van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: Findings from 2 citizens' juries. *Journal of the American Medical Informatics Association: JAMIA, 28*(10), 2128–2138. https://doi.org/10.1093/jamia/ocab127

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Walter, S. D., Feinstein, A. R., & Wells, C. K. (1987). Coding ordinal independent variables in multiple regression analyses. *American Journal of Epidemiology, 125*(2), 319–323. https://doi.org/10.1093/oxfordjournals.aje.a114532

Woodruff, A., Anderson, Y. A., Armstrong, K. J., Gkiza, M., Jennings, J., Moessner, C., Viegas, F., Wattenberg, M., Webb, L., Wrede, F., & Kelley, P. G. (2020). "A cold, technical decision-maker": Can AI provide explainability, negotiability, and humanity? arXiv:2012.00874 [Cs]. http://arxiv.org/abs/2012.00874.

Zhang, Y., & Mittal, V. (2005). Decision difficulty: Effects of procedural and outcome accountability. *Journal of Consumer Research, 32*(3), 465–472. https://doi.org/10.1086/497558