"It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems

Saumya Pareek
The University of Melbourne
Melbourne, VIC, Australia
spareek@student.unimelb.edu.au

Sarah Schömbs
The University of Melbourne
Melbourne, VIC, Australia
sschombs@student.unimelb.edu.au

Eduardo Velloso
The University of Sydney
Sydney, NSW, Australia
eduardo.velloso@sydney.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, VIC, Australia
jorge.goncalves@unimelb.edu.au

Abstract

Humans naturally seek to identify causes behind outcomes through causal attribution, yet Human-AI research often overlooks how users perceive *causality behind AI decisions*. We examine how this perceived locus of causality—internal or external to the AI—influences trust, and how decision stakes and outcome favourability moderate this relationship. Participants (N=192) engaged with AI-based decision-making scenarios operationalising varying loci of causality, stakes, and favourability, evaluating their trust in each AI. We find that internal attributions foster lower trust as participants perceive the AI to have high autonomy and decision-making responsibility. Conversely, external attributions portray the AI as merely "a tool" processing data, reducing its perceived agency and distributing responsibility, thereby boosting trust. Moreover, stakes moderate this relationship—external attributions foster even more trust in lower-risk, low-stakes scenarios. Our findings establish causal attribution as a crucial yet underexplored determinant of trust in AI, highlighting the importance of accounting for it when researching trust dynamics.

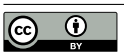
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

Keywords

causal attribution, locus of causality, trust, human-AI interaction, human-AI decision-making, decision stakes, outcome favourability

ACM Reference Format:
Saumya Pareek, Sarah Schömbs, Eduardo Velloso, and Jorge Goncalves. 2025. "It's Not the AI's Fault Because It Relies Purely on Data": How Causal Attributions of AI Decisions Shape Trust in AI Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713468



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713468

1 Introduction

Consider an individual, John, bursting into laughter at a comedian's joke. In this scenario, one could attribute the outcome—John's laughter—to the person himself (John) or the stimulus (the comedian's joke). If John is the only one laughing and tends to laugh easily at any comedian, then one would attribute the cause behind his laughter *internally* to him, to his sense of humour. Alternatively, if John, usually reserved at comedy shows, suddenly laughs alongside the entire audience, then one might believe the comedian's exceptional skills brought about John's laughter, attributing the cause *externally* to John. Both scenarios involve the same outcome—John's laughter—but our understanding of the outcome differs based on where we attribute its causality [23]. **Considering how perceptions of causality influence our understanding of everyday events, how might they impact our interactions with and trust in AI systems?**

Attribution theory posits that humans instinctively seek to understand the causes behind actions and outcomes [33]. This innate drive for causal understanding, termed *causal attribution*, influences how we attribute responsibility and, subsequently, how we assign blame or praise [48, 57]. An intelligent agent could cause an outcome, yet individuals could perceive its causal role differently. For example, if a self-driving car brakes suddenly without an apparent obstacle, some users might attribute the cause to an *internal* issue (e.g., limitations in the car's decision-making software), which could decrease their trust in the system. In contrast, others may attribute the error to *external* causes (e.g., poor road markings or inclement weather conditions), potentially maintaining their trust in the system's capabilities. This presents an opportunity to empirically examine how variations in the perceived locus of causality behind automated decisions—whether seen as stemming from causes *internal* or *external* to AI—impact trust in the AI.

Notably, the integration of AI-based decision-making in everyday life is driven by the recognition that its adoption can reduce costs, enhance performance, and facilitate more objective decision-making. However, AI systems can make mistakes, so end-users must discern when to trust their output [16, 37]. In response, research efforts have focused on identifying the factors that influence user trust in AI, aiming to appropriately calibrate this trust to match system capabilities [22, 45, 53, 66]. However, existing approaches to trust calibration—often revolving around providing explanations [50, 51]

or confidence scores [58, 83]—encounter critical challenges. Explanations can paradoxically promote over-reliance by creating a false sense of legitimacy [17, 52], while stated confidence scores are often overshadowed by the performance of the AI observed in practice [80, 81]. While these methods aim to increase transparency, they fail to account for how individuals perceive and interpret the *causal basis* behind AI decisions—whether they perceive the cause behind an AI outcome as originating from within the AI itself (attributing it to the AI's algorithms and decision-making processes) or from external factors (such as the data the AI processes or the level of human oversight) [70, 77]. Tomlinson and Mayer [70], in their trust calibration model, particularly theorise that trust may decline more sharply when negative AI outcomes are attributed internally to the AI. However, the impact of causal attributions on human-AI trust dynamics remains to be empirically examined. Understanding this relationship could also offer valuable insights for refining trust calibration approaches, ensuring better alignment with user expectations and attribution behaviours.

Further, individuals are more likely to trust and accept favourable AI decisions, a tendency known as *outcome favourability bias* [10, 44]. Outcome favourability modulates trust in AI based on the advantage or disadvantage resulting from the decision [75], but differing causal attributions may further complicate this relationship. For example, when the cause behind an unfavourable outcome is attributed *internally* to the AI—i.e. to its decision-making process or capabilities—trust may be impacted differently compared to when an unfavourable outcome is attributed to factors *external* to the AI, outside its locus of control. This raises the question: *is trust in AI shielded from the consequences of poor decision-making when individuals perceive an external locus of causality?* Additionally, individuals are more inclined to trust and accept AI decisions in low-stakes contexts such as music recommendation [63], compared to high-stakes contexts such as medical diagnostics [31] where the consequences of decisions are substantial. However, it remains unclear how trust is influenced by the consequences or risks of the decision-making context when the locus of causality behind decisions is also manipulated.

Therefore, this work aims to bridge the aforementioned gaps by systematically investigating how the attribution of causality shapes trust in AI systems and how decision stakes and outcome favourability impact this relationship. We aim to answer the following research questions:

- **RQ1:** How does the perceived locus of causality behind AI decisions—attributed internally or externally to the AI—influence trust in the AI?
- **RQ2:** How do contextual factors, such as the stakes of the decision-making context and the favourability of the AI outcome, moderate the impact of causal attributions on trust in AI?

We conducted a 2 (*causal attribution: internal vs. external)* × 2 (*decision stakes*: high vs. low) × 2 (*outcome favourability*: favourable vs. unfavourable) within-subjects scenario-based experiment with 192 participants. Scenarios introduced a high- or low-stakes decision-making context in which an AI operated, subsequently revealed the AI's decision, and signalled a specific locus of causality behind this decision—operationalising our three independent variables. The locus of causality was operationalised using Kelley's framework of causal attribution [32], which outlines three information variables: *consensus* (whether the AI's decision aligns with those of other systems), *distinctiveness* (whether the decision is specific to the given input), and *consistency* (whether the decision remains stable over time and after repeated exposure to the same input). *Internal* attributions signalled that the AI's decision stemmed from its inherent algorithms and capabilities (low consensus, low distinctiveness, and high consistency), while *external* attributions highlighted the influence of external factors, such as data quality or environmental factors (high consensus, high distinctiveness, and high consistency). A manipulation check confirmed that our scenarios robustly operationalised our manipulations as intended. For each scenario, after learning about the AI's decision and an associated locus of causality, participants reported their *Situational Trust* in the AI. To deepen our understanding of trust dynamics, participants also answered open-ended questions exploring the factors influencing their trust in each scenario.

Our findings demonstrate a critical relationship between participants' perceptions of causality and their trust in AI systems. When participants attribute the locus of causality behind an AI's decision internally to the AI, their trust is notably lower. In such cases, participants perceive the AI as more autonomous and responsible, having full control and agency in the decision-making process. This perception raises concerns about the AI's excessive autonomy and insufficient human oversight, consequently reducing participants' trust. Conversely, when participants attribute causality externally to an AI, they trust the AI substantially more. In such cases, participants distribute responsibility among various entities and stakeholders in the decision-making ecosystem, and perceive the AI more as *"a tool"* that processes data, operating under the influence of factors beyond its control, such as the quality of data supplied to it. This external attribution of causality reduces perceived AI agency and reassures participants about the AI's role in the decision-making process, thereby enhancing their trust.

Additionally, we observe that the effect of causal attribution on trust varies with the decision stakes. While external attributions generally foster higher trust, this effect is more pronounced in low-stakes scenarios, where the perceived risk and consequences of trusting the AI are less severe compared to high-stakes scenarios. Furthermore, our analysis reveals no significant interaction between outcome favourability and causal attribution, showing no evidence of a differential effect of causal attribution on trust across favourable and unfavourable decisions. This highlights the role of causal attribution as an important, stable determinant of trust. Lastly, our findings reiterate the impact of decision stakes and outcome favourability on trust: participants exhibited higher trust in low-stakes scenarios and favourable decisions than in high-stakes scenarios and unfavourable decisions.

Our work makes the following contributions. First, we identify causal attribution as a critical yet previously overlooked determinant of trust in AI. We show that trust varies based on whether users perceive decisions as stemming from the AI's own capability and algorithms, or as reliant on external factors such as data quality and stakeholders. Second, we uncover how causal attributions shape perceptions of AI agency, autonomy, responsibility, and the extent of human oversight in decision-making. Through several interesting qualitative insights, we discuss how different loci of causality lead participants to hold *the same AI* differently responsible for its decisions, ultimately impacting the scrutiny extended to and trust placed

in it. We advocate for framing AI decisions to clearly convey causality and shared responsibility as an effective method to boost transparency and foster (appropriate) trust in AI. Third, by demonstrating how attributing excessive agency to AI can lead to decreased trust and increased scepticism, our work emphasises the need to shift the discourse away from portraying AI systems as overly autonomous and agentic. Fourth, we highlight how the impact of causal attribution on trust varies with decision stakes—underscoring the need for contextual trust-building strategies. While there is no silver bullet for fostering trust in high-stakes environments even when AI performance warrants it, explicating causality can be effective, especially when complemented by other transparency-boosting strategies.

Ultimately, our results demonstrate that users' trust in AI is sensitive to their understanding of the causal mechanisms behind AI decisions, and advocate for their transparent communication. We underscore the need for future trust calibration efforts to consider *where* users attribute causality and account for these perceptions when studying trust dynamics.

2 Related Work

2.1 Trust in AI Systems and its Determinants

In this work, we adopt the definition of trust proposed by Lee and See [37], who describe it as *"an attitude that an agent will achieve an individual's goal in a situation characterised by uncertainty and vulnerability."* Stemming from this notion, multiple definitions and decompositions of human-AI trust have been proposed, all converging toward the central elements of *uncertainty*, *vulnerability*, and *expectations* [72]. Trust, therefore, emerges as a dynamic, temporal characteristic of any human-AI interaction fraught with uncertainty and vulnerability. As AI agents become increasingly intertwined with our everyday lives, aligning user trust with the capabilities and limitations of such AI agents becomes crucial. Hoff and Bashir [24] further classify trust in automated systems into distinct types, including *dispositional* and *situational* trust.

Dispositional trust refers to an individual's inherent tendency to (dis)trust automation in general, while *situational trust* is shaped by the specific contextual factors surrounding human-AI interaction, such as task complexity, AI performance, and the perceived risk associated with the decision-making process [24]. Importantly, different levels of dispositional trust can result in both *automation bias*, where individuals place unwarranted trust in automated advice because of the perception that automation is superior [19, 49], or *algorithm aversion*, where users are sceptical of automated advice and disregard it even when it may be reliable [30, 56]. Thus, while *dispositional trust* might predispose individuals towards a certain level of trust in AI, *situational trust* adjusts based on the context of each interaction with a particular AI. In this work, we investigate and consider both *dispositional* and *situational trust*, aligning with Hoff and Bashir's call for research that examines the various layers of end-users' trust in automation to appropriately understand how trust is modified through interactions with AI systems [24].

2.1.1 Determinants of Trust. Research has explored methods to calibrate trust in AI agents, employing both 'endo' (during the interaction) and 'exo' (before or after the interaction) techniques, as well as static and adaptive approaches (see Wischnewski et al. [78] for a comprehensive overview). Trust calibration approaches typically centre around the various determinants of trust identified so far [36]. These determinants pertain to the decision-making *process* or *performance* of an AI [78], communicated through explanations [50, 51] and confidence scores respectively [58, 83]. However, despite their promise, these approaches encounter critical challenges. Explanations can backfire and create a false sense of legitimacy, inadvertently promoting over-reliance on AI [17, 52]. Moreover, the influence of stated confidence scores on trust can be overshadowed by the AI accuracy or behaviour observed in practice [80, 81]. A growing body of research underscores the importance of perceived responsibility and accountability in shaping user trust in AI systems [6, 12, 68]. When users perceive probabilistic AI systems as autonomous entities capable of making decisions, questions around perceived responsibility and trust become increasingly complex [20, 26, 69]. Who do users hold responsible for an AI's errors—the system itself or its human designers? Such individual responsibility ascriptions may also directly influence trust in many ways, as perceptions of responsibility can affect users' willingness to rely on AI systems. For example, Robinette et al. [60] find that in emergency scenarios, users who attributed a robot's decisions to the (assumed) competence and accountability of its human creators were willing to follow the robot into increasingly dangerous situations. In fact, users' trust is sensitive to even subtle indicators of agency (and indirectly, causality): framing AI systems as 'intelligent' or 'autonomous' agents can significantly shape perceptions of their trustworthiness, albeit to different extents based on individual and contextual factors [26, 27]. However, the impact of perceived causality on trust remains to be systematically examined. Much of the current research on human-AI trust overlooks *where users perceive the causality behind AI decisions to lie*, often failing to account for whether users think "the AI" itself is making decisions or if decisions are influenced by factors outside its locus of control.

Attribution theory suggests that individuals naturally tend to seek the causes behind actions and outcomes [33]. This drive for causal understanding informs how they assign responsibility and distribute blame or praise [48, 57]. Importantly, at the core of responsibility attribution lies the human ability to construct causal narratives that explain others' actions [33]. Causality, thus, is a core concept that governs judgements of responsibility [18]. Responsibility attribution, then, becomes a proximal consequence of causal attribution. An intelligent agent could cause an outcome, yet individuals could perceive its causal role in the process differently. This raises the question: **would individuals perceiving different loci of causality behind AI decisions exhibit different levels of trust in the same AI?** Can perceptions of causality be a significant, yet previously overlooked, determinant of trust? Moreover, current approaches to foster appropriate trust in AI, such as explanations or confidence scores, focus primarily on observable system performance or process but overlook deeper user perceptions about the causal mechanisms behind AI decisions [24, 37]. Thus, understanding how causal attributions shape trust in AI could offer valuable insights to refine these approaches, aligning them with individual user perceptions of causality.

2.2 Attributing Causality Behind Outcomes

Weiner's causal attribution theory [77] explains how individuals make sense of their experiences by attributing causes to behaviours

and events. He posits that individuals respond to outcomes, especially negative ones, by identifying their causes (i.e., attributing their causality to *something*) and then evaluating these causes along three dimensions: ***locus of causality*** (whether the cause is internal or external to the agent), *controllability* (the degree of control the agent has over the outcome), and *stability* (whether the cause is constant or variable). Of particular interest to this work is the locus of causality dimension. An individual could attribute the cause of an AI's decision as either being ***internal*** to the AI (i.e., to factors pertaining to the automation itself, such as its programming logic) or ***external*** to the AI (i.e., to situational factors, such as the quality of the data it is trained on).

Building upon Weiner's theory, Tomlinson and Mayer [70] introduced a causal attribution model specific to trust repair in interpersonal relationships. This model integrates Weiner's attribution dimensions with the trust feedback loop from Mayer et al. [47], illustrating how trust evolves through a cyclic process of risk-taking, outcome evaluation, and trust adjustment based on perceived trustworthiness. Trustworthiness itself is assessed through the lens of an agent's *ability*, *benevolence*, and *integrity*, with these perceptions directly influencing trust levels. According to Tomlinson and Mayer's [70] model, if end-users believe that a negative decision by an AI is due to the AI's own capabilities, i.e., they ascribe an *internal* causality to the outcome, their trust in the AI is likely to decrease. Conversely, if the cause of the negative decision is perceived as *external* to the AI, their trust in the system may not necessarily diminish. **This presents a novel opportunity to examine the extent to which users' trust in an AI is tied to the perceived locus of causality behind its decisions**.

Kelley [32] outlined three information variables that affect how people assign causality—namely, ***consensus***, ***distinctiveness*** and ***consistency***. This information model allows three possible ways of attribution, based on the factor that caused a variation in some effects: (a) over decision-making entities (from which *consensus* is derived); (b) over stimuli (from which *distinctiveness* is derived); and (c) over time/interactions (from which *consistency* is derived). More specifically, ***consensus*** pertains to whether a response aligns with that of others facing the same stimulus, ***distinctiveness*** refers to whether a response is associated distinctively with the stimulus, and ***consistency*** deals with whether the response is consistent over time and after multiple exposures to the same stimulus. These three variables thus signal the attribution of causality behind a decision as follows:

- **Internal Attribution:** Users are more likely to attribute the cause as internal to the decision-making entity when presented with information that indicates low *consensus*, low *distinctiveness*, and high *consistency* [23, 32]. Consider the scenario where a student, Alex, fails a maths exam. In this scenario, the outcome—failing the exam—could be attributed to Alex themself or the stimulus (the exam). If most classmates passed the exam while Alex did not (low *consensus*), Alex also struggles in other subjects (low *distinctiveness*), and if Alex consistently found the maths exam difficult across multiple attempts (high *consistency*), then the poor performance is likely attributed to factors internal to Alex, such as their insufficient preparation or understanding of the subject matter.
- **External Attribution:** Users are more likely to attribute the cause as external to the decision-making entity when presented with information that indicates high *consensus*, high *distinctiveness*, and high *consistency*. Consider another version of the scenario where Alex fails a maths exam. If most classmates also failed the exam (high *consensus*), Alex generally excels in other subjects (high *distinctiveness*), and if Alex consistently found the exam difficult across multiple attempts (high *consistency*), then the poor performance is likely attributed to external factors, such as the intrinsic difficulty of the exam or insufficient teaching, rather than internally to Alex.

In this work, we utilise the aforementioned three information variables to manipulate the causal attribution behind AI decisions to be either *internal* or *external* to the AI.

2.3 Contextual Factors affecting Perceptions of Trust in Automated Decision-Making

Factors related to the decision-making context play a crucial role in shaping users' trust in AI. Firstly, the **favourability** of an automated outcome impacts users' trust and fairness perceptions [75]. Individuals are more inclined to trust and accept AI decisions when the outcomes align with their interests, a tendency known as *outcome favourability bias* [10]. This bias can complicate trust appropriateness, as it may overshadow the objective evaluation of an AI system's trustworthiness, particularly when it provides an unfavourable outcome. While outcome favourability bias influences trust based on the direct benefit or loss experienced by users, varying causal attributions can further complicate this relationship. Building on Tomlinson and Mayer's causal attribution model [70], which theorises that trust may decline more sharply when unfavourable outcomes are attributed *internally* to the AI, we hypothesise that outcome favourability may moderate the relationship between causal attributions and trust. In particular, when an unfavourable outcome is attributed internally to the AI—perceived as a failure of the system's capability or decision-making process—trust may decrease more sharply compared to when the unfavourable outcome is attributed to external factors beyond the AI's control. However, it remains to be empirically examined how internal and external loci of causality behind AI decisions differently influence trust, when these decisions are favourable or unfavourable. In this work, we aim to bridge this gap.

Secondly, trust in AI systems is influenced by the **stakes** of the decision-making context [1, 31], which pertain to how grave the consequences of a decision can be. In high-stakes domains such as hiring [35, 40], medical diagnostics [11, 39, 50] and criminal justice [14, 43], the outcomes of AI decisions carry significant implications for individuals' lives. Conversely, in lower-stakes domains such as personalised shopping [42] and music recommendations [63], the consequences of AI decisions are less severe, often only violating personal preferences at their worst. AI systems involved in low-stakes decision-making are generally perceived as more trustworthy than those in high-stakes situations, where users are more sensitive to perceived risks [2]. This *heightened* risk sensitivity in high-stakes contexts suggests that perceived decision stakes may intensify the impact of causal attributions on trust. Specifically, as the potential risks or consequences associated with an AI's decision increase, the role of causality may become more critical in shaping users' trust.

This warrants further exploration into how stakes might moderate the relationship between causal attribution and trust, a gap we address in this work.

Pop et al. [55] conducted a preliminary investigation into how (solely) an internal causal attribution impacted the perceived reliability of an agent, finding that internal attributions make individuals with a higher dispositional trust in automation less sensitive to changes in AI accuracy. Our study builds upon their work by exploring both internal and external causal attributions and their impact on trust. This broader approach addresses a crucial gap by investigating how trust is shaped not only by factors intrinsic to the AI but also by those external to its locus of control. Another significant limitation of existing research is that it does not consider the interplay of causal attributions with decision stakes and outcome favourability. These factors are intrinsically intertwined in real-world AI-based decision-making, and their collective influence on trust dynamics can differ markedly from their individual impacts.

Therefore, in this work, we address the above gaps by systematically evaluating the effects of causal attribution, decision stakes, and outcome favourability on trust. Our goal is to examine how each factor individually influences trust perceptions across scenarios, while also examining whether the impact of causal attributions on trust is moderated by stakes and favourability. Additionally, we investigate how dispositional factors, such as an individual's general propensity to trust automation [34], shape trust in this experimental context.

3 Method

To explore how *causal attributions*, *decision stakes*, and *outcome favourability* impact trust, we conducted a within-subjects survey-based experiment employing a mixed-methods design. This approach allowed us to collect both quantitative scores of trust perceptions and qualitative insights into the reasons behind these perceptions. In the following sections, we present the design considerations for crafting scenarios, discuss experimental manipulations, report manipulation check findings, and describe the main experiment procedure.

3.1 Scenario Selection and Design

We adopted a scenario-based approach which is widely employed in HCI research to elicit user opinions, attitudes, and trust in a controlled manner [2, 7, 31, 38, 62]. This choice is also supported by the finding that participants' behaviours during scenario-based studies often mirror their real-world reactions and decision-making processes [38, 79].

3.1.1 Operationalising Causal Attributions, Stakes, and Outcome Favourability. We chose four scenario contexts commonly involving AI-based decision-making, and designed scenario variants for each combination of *decision stakes*, *outcome favourability*, and *causal attributions*. We operationalised *stakes* through the high and low severity or risk posed by the scenario context—our two *high stakes* scenarios comprised (1) medical diagnostics and (2) hiring decisions, while the two *low stakes* scenarios involved (3) music recommendations and (4) weather-based clothing recommendations [31, 62]. Further, we manipulated *outcome favourability* by designing AI decisions that have a positive (*favourable*) or negative (*unfavourable*) impact on the human, such as approving a candidate to be hired for a job versus rejecting them. Lastly, *causal attributions* were signalled using Kelley's framework of three information variables: *consensus*, *distinctiveness*, and *consistency*, that govern how we ascribe causality. The *high* and *low* values assigned to these variables follow Kelley's foundational definitions [32], and are further validated by additional research [23, 55]. An *internal* attribution indicated that decisions stemmed from the AI's inherent capabilities and algorithms, whereas an *external* attribution suggested decisions were dependent on external factors, such as the quality of data supplied by external entities (e.g., the meteorological department in scenarios involving our Weather AI). Table 1 illustrates how we operationalised causality by detailing how each variable was expressed in the scenarios and providing example sentences from the high-stakes medical diagnostics scenario. We note that, since *consistency* represents the stability of an AI's decision-making across similar cases over time, it is held high and constant across both attributions to emphasise this stability, and to convey that the locus of causality remains constant and unchanging (i.e., it does not shift between *internal* and *external* within an interaction) [32].

3.1.2 Scenario Design and Creation. We systematically fixed both the participant's and the AI's role across scenarios to eliminate any confounding influences and ensure that the observed participant behaviours were attributable solely to our experimental manipulations. The roles are as follows:

- Participants always assume the role of an actor within the scenario, being directly impacted by the AI's decision. This design choice follows similar work [62] and aims to minimise the potential influence of the actor-observer bias, wherein individuals tend to attribute their own actions to situational *(external)* factors and others' actions to their personal *(internal)* traits [29].
- The AI always functions as the decision-maker [21]. This design choice eliminates any potential confounding effects that might stem from shared responsibility in collaborative decision-making between the participant and the AI [41, 82].

We structured each scenario as follows: first, the decision-making context and the AI were introduced, highlighting how the human would be subject to the AI's decision, operationalising our independent variable *decision stakes*. Subsequently, the AI's decision was revealed, operationalising our second independent variable *outcome favourability*, and then a specific locus of causality behind this decision was signalled, operationalising our third independent variable *causal attribution*. We utilised two levels each for *decision stakes* (high or low), *outcome favourability* (favourable or unfavourable), and *causal attribution* (internal or external), with two distinct scenarios for each stake level, necessitating the generation of 16 unique scenarios to cover all variable combinations.

To generate the 16 scenario texts, we utilised ChatGPT (GPT-4), a large language model trained by OpenAI [1]. We iteratively refined our prompts to specify desired scenario characteristics, definitions of independent variables, experimental manipulations, and additional contextual details. The complete prompt and the final scenario texts are included in Appendix A. We reviewed the generated scenario

[1] https://openai.com/gpt-4

Table 1: Kelley's three information variables (*consensus, distinctiveness,* and *consistency*) that influence how we perceive causality behind decisions, their interpretation, and operationalisation in our scenario texts.

Locus of causality	Kelley's information variables and high/low values, collectively signalling a locus of causality [32]	An example operationalisation of these variables, in the medical diagnostics scenario
Internal to the AI	↓ **Low Consensus**: The AI's decisions frequently differ from those made by other similar AI systems analysing the same input data.	"MediScan AI frequently provides diagnoses that differ from those given by other diagnostic AIs for similar patient data."
	↓ **Low Distinctiveness**: The AI's behaviour is not distinct, i.e., it remains consistent across different situations, suggesting that the behaviour is a characteristic of the AI itself rather than being influenced by situational factors.	"MediScan AI's detection performance remains the same irrespective of the patient data it is assessing."
	↑ **High Consistency**: The AI consistently makes the same decisions when presented with the same input, highlighting stable and predictable behaviour.	"When analysing the same set of patient records repeatedly, MediScan AI provides the same diagnosis."
External to the AI	↑ **High Consensus**: The AI's decisions align with those made by other similar AI systems analysing the same input data.	"MediScan AI's diagnoses are frequently consistent with those given by other diagnostic AIs for similar patient data."
	↑ **High Distinctiveness**: The AI's behaviour changes significantly with different inputs or conditions, suggesting that its decisions are largely influenced by the specifics of the current situation rather than its inherent attributes.	''MediScan AI's detection performance only varies depending on the patient scans and data the clinic supplies to it."
	↑ **High Consistency**: The AI consistently makes the same decisions when presented with the same input, highlighting stable and predictable behaviour.	"When analysing the same set of patient records repeatedly, MediScan AI provides the same diagnosis."

texts to ensure alignment with our instructions and study requirements. Utilising a language model to generate scenario texts allowed us to maintain consistency across scenarios, minimising unintended variability and ensuring variations in participant responses can be attributed to our manipulations rather than random differences in stimulus wording, following past research [8, 15].

3.2 Manipulation Check

3.2.1 Method. We conducted a series of pilot tests and a manipulation check to ensure the generated scenarios accurately operationalised our independent variables and our manipulations were perceived as intended. Feedback from the pilot testing helped us disambiguate confusing sentences and enhance clarity. For the manipulation check, each participant was presented with four of the 16 scenarios, each featuring a distinct combination of our independent variables. This selection was structured to ensure all participants equally experienced *internal* and *external* causal attributions, *high* and *low* stakes, and *favourable* and *unfavourable* outcomes. We randomised the order in which we presented the scenarios to control for ordering effects. Participants were asked to read the scenario, imagine themselves in the given situation, and report their perceptions of each AI decision's stakes and locus of causality. Outcome favourability was not subjected to a manipulation check because the favourability of the AI's decision was explicit within scenario texts.

3.2.2 Measures and Participants. For perceived stakes, participants rated the significance of the consequences of the AI's decision on a 4-point scale ranging from 1 ("Not significant at all") to 4 ("Very significant"), following past research [2, 62]. For the locus of causality, we adopted Russell's causal dimension scale [61], specifically the *'locus of causality'* sub-scale, which consisted of three items rated on 9-point semantic differential scales. Consequently, the locus of causality scores ranged from 3 to 27, with higher scores indicating participants attributed the decision *internally* to the AI, and lower scores suggesting attribution to factors *external* to the AI's locus of control.

We recruited 24 participants who were located in the United States, were native English speakers, and had a platform approval rating $\geq 98\%$, through the crowdsourcing platform Prolific [2]. Participants took a median time of 6 minutes to complete the survey and received US$2 for participation.

3.2.3 Results. A Wilcoxon signed-rank test revealed a statistically significant difference in how participants perceived the impact of decisions between *high* ($Median = 4$, $M = 3.68$, $SD = 0.46$) and *low* ($Median = 2$, $M = 2.35$, $SD = 0.83$) stake scenarios ($V = 850.5$, $p < .001$). This finding validates our stakes manipulation; *high* stakes scenarios were indeed regarded by participants as considerably more consequential than their *low* stakes counterparts (Figure 1 (left)). Further, we conducted a paired t-test to verify the effectiveness of our causal attribution manipulation. Results showed a statistically significant difference in the perceived locus of causality between scenarios designed with *internal* attributions ($M = 22.16$, $SD = 4.93$) and those with *external* attributions ($M = 12.29$, $SD = 6.50$) ($t(47) = 7.54$, $p < .001$, 95% CI between 7.24 and 12.51). Since higher scores on the causal scale indicate an *internal* attribution, these findings validate our causal attribution manipulation: scenarios intended to portray a locus of causality internal (external) to the AI were indeed perceived as such by participants (Figure 1 (right)).

Given these findings, we are confident that our scenarios effectively operationalised our intended experimental manipulations and were thus suitable for the main experiment. The full scenario texts are included in Appendix B.

[2] https://www.prolific.com/

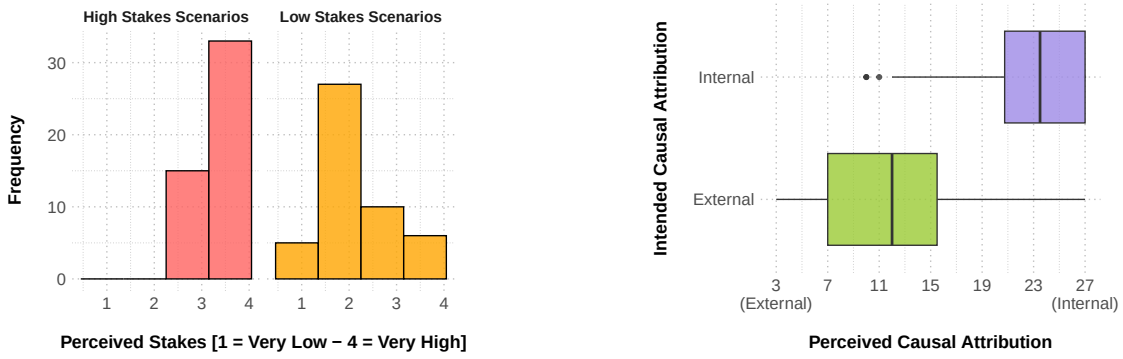


Figure 1: Manipulation check outcomes. (left): Comparison of perceived stakes (1 = Very Low, 4 = Very High) against the stakes operationalised in the scenarios. (right): Comparison of perceived causal attributions (lower scores = external attribution, higher scores = internal attribution) against attribution intended through the scenarios.

3.3 Main Experiment

The main experiment employed a 2 (*causal attributions*: internal vs. external) × 2 (*decision stakes*: high vs. low) × 2 (*outcome favourability*: favourable vs. unfavourable) within-subjects factorial design. We deliberately did not introduce a condition where causality was not signalled. Past research suggests that in the context of AI systems, the absence of causal information is rarely 'neutral' and can lead to assumptions or default attributions based on individual biases or past experiences [55], which could confound the causal impact we aim to measure.

In addition to testing the direct influence of these three binary predictors on situational trust in AI, we also sought to investigate how the impact of causal attribution is moderated by the *stakes* and the *favourability* of the decision. Thus, we examined two interaction effects, each exploring the interplay between *attribution* and *stakes*, and *attribution* and *favourability*. We did not examine a three-way interaction (*stakes*, *favourability*, and *attribution*) as no strong theoretical or empirical basis in the literature supports a joint moderation effect of both favourability and stakes on how causal attribution impacts trust in human-AI interaction. To calculate our sample size, we utilised the R package `InteractionPoweR` [4], which accounts for the larger sample size requirements of interaction effects and allows variables to be non-continuous (i.e., binary or ordinal) [5], both requirements of our experimental setup. The minimum recommended sample size was 182 participants, considering an $\alpha = 0.05$, and a power of 0.8 [13, 67, 71]. To ensure balance across our experimental conditions, we conservatively recruited 192 participants, with a mean age of 40.19 years ($SD = 14.35$).

We deployed our study on Prolific, utilising the same participant screening criteria as our manipulation check, and recruited an equal number of men and women who had not participated in our manipulation check. Participants joined our study exactly once, and passed at least one of two attention checks, thus no data was excluded from analysis. Participants took a median time of around 18 minutes to complete the survey and were compensated US$4.70 for their time, well above the minimum hourly wage recommended by Prolific [3]. Our university's Human Ethics Committee approved the study.

[3]https://researcher-help.prolific.com/en/article/9cd998

3.3.1 Measures. Dispositional Trust: We captured participants' dispositional trust in automation using the Trust in Automation - Propensity to Trust (TiA-PtT) questionnaire [34] (Fig 2 (a)), as dispositional trust is known to influence trust experienced in AI systems [52, 53, 74]. This measure captures an individual's general propensity or inclination to trust automation, irrespective of the specific context or type of technology.

Situational Trust: Captured after the AI's decision and the associated causal attribution are revealed, *Situational Trust* reflects the actual trust a participant experiences in the AI post-decision [24]. To measure this, we employed the TXAI scale [25], which has been validated specifically for use in Human-AI contexts by Perrig et al. [54]. Following their recommendations, we excluded the potentially problematic scale items they identified, ensuring the use of a robust and reliable instrument to measure trust across our scenarios. The resultant 4-item TXAI questionnaire was administered on a 5-point Likert scale ranging from 1 ("Strongly disagree") to 5 ("Strongly agree"). In our study, the TXAI scale demonstrated high internal reliability with a Cronbach's α of 0.95 (95% CI between 0.94 and 0.96).

3.3.2 Procedure. Fig 2 illustrates the complete experiment flow. Each participant was assigned one level of *decision stakes*: high vs. low, *outcome favourability*: favourable vs. unfavourable, and *causal attributions*: internal vs. external for every scenario (Fig 2 (b, c, d)). Each participant engaged with four scenarios, selected in a strategically counterbalanced way to ensure every level of our independent variable was encountered an equal number of times by every participant. This counterbalancing controlled for potential order effects and provided a balanced representation of conditions across the participant pool.

The survey began with a pre-task questionnaire (Fig 2 (a)) to gather participants' demographic information and their *Dispositional Trust* (TiA-PtT). We then informed participants that they would be presented with a series of scenarios involving AI-based decision-making, instructing them to imagine themselves in these contexts and answer questions that followed.

Participants read each scenario, which introduced the decision-making context, subsequently revealing the AI decision and signalling a locus of causality behind it (Fig 2 (e)). Participants were

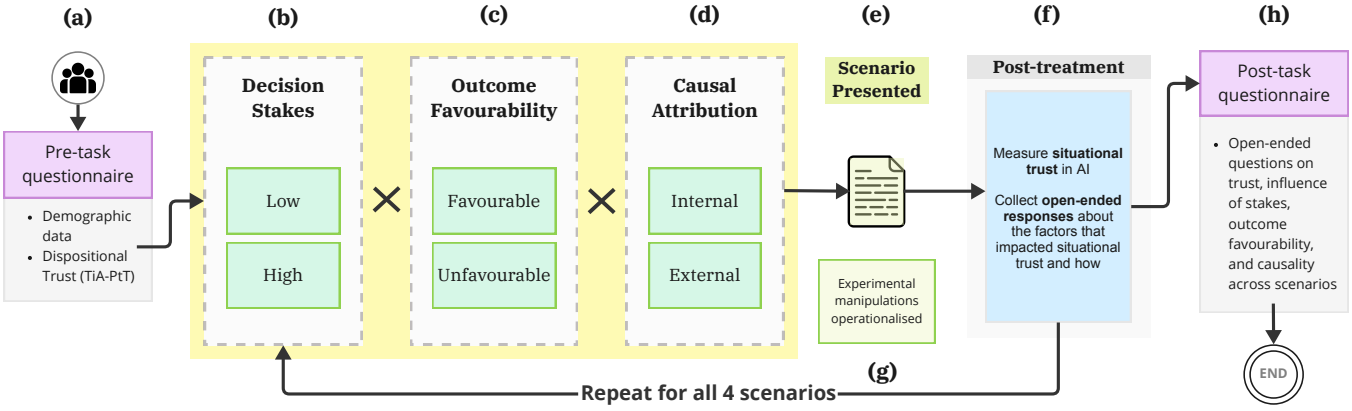


Figure 2: The full experiment flow. All participants view four scenarios in a randomised manner, with combinations of stakes, outcome favourability, and causal attribution strategically counterbalanced across participants. (a): Pre-task questionnaire, demographic data and *Dispositional Trust* measured. (b, c, d): For each scenario, a combination of decision-making stakes (*low* or *high*), outcome favourability (*favourable* or *unfavourable*), and causality behind the decision (*internal* or *external* to the AI) chosen for each participant. (e): Corresponding scenario text presented, describing the decision-making context, subsequently revealing the AI's decision, and signalling a locus of causality behind it. (f): Participants' *Situational Trust* in the AI measured. (g): Process repeated for all four scenarios seen by a participant. (h): Post-task questionnaire and debriefing.

then asked to report perceptions of their *Situational Trust* in the given AI (Fig 2 (f)). These Likert scales included an attention-check question at random, asking participants to select a specific scale response. Each participant was presented with two attention-check questions.

Following each scenario, participants responded to two open-ended questions specifically exploring the factors that influenced their trust in the given AI. At the study's conclusion, after all four scenarios were shown, two more open-ended questions were asked, which encouraged participants to reflect on their trust perceptions across the AIs in the different scenarios and describe how they perceived causes behind these automated decisions (Fig 2 (h)).

4 Results

We employ a Cumulative Link Mixed Model (CLMM) to investigate how the perceived locus of causality behind AI decisions (RQ1) and contextual factors such as the decision stakes and outcome favourability (RQ2) influence *Situational Trust* in AI. We further perform post-hoc analyses to obtain pairwise contrasts between different levels of our independent variables and report the corresponding Estimated Marginal Means (EMM). Details of participant demographics are presented in Appendix C.

4.1 Model Construction

Participants' *Situational Trust* in AI (Section 3.3.1) formed our dependent variable. The scale comprised four 5-point Likert scale items, each ranging from 1 (indicative of low perceived trust) to 5 (indicative of high perceived trust). Given their ordinal nature, we employed a Cumulative Link Mixed Model (CLMM) to investigate the effects of the independent variables *stakes*, *favourability*, and *attribution* on *Situational Trust*. In this model, we also included *Dispositional Trust* (TiA-PtT) to account for how participants' general disposition to trust automation may impact their *Situational Trust*. Despite the high internal consistency and reliability of our trust scale (Cronbach's α of 0.95, 95% CI between 0.94 and 0.96), we included question IDs (QID) of scale items as random effects in our model to account for any potential variability in responses that could stem from specific scale items. Additionally, participant IDs (PID) were incorporated as random effects to control for individual differences and potential correlations amongst repeated measurements from the same participant. The resultant CLMM function was as follows: `Situational_Trust ~ TiA-PtT + Stakes + Favourability + Attribution + Attribution:Stakes + Attribution:Favourability + (1|QID) + (1|PID).`

We employed the statistical R package `ordinal` to build our CLMM. We calculated the Variance Inflation factor (VIF) to check for multicollinearity across the independent variables, and the obtained VIF values ranged from 1.00 to 1.51, well below the commonly used threshold of 5 to detect multicollinearity [59].

4.2 Quantitative Results

The results of our CLMM analysis are presented in Table 2, and Estimated Marginal Means (EMMs) obtained from post-hoc analyses are illustrated in Figure 3.

We observed a statistically significant main effect of **stakes** on *Situational Trust* (β = 1.510, SE = 0.102, $p < 0.001$). Participants were more likely to trust AI in *low* stakes scenarios (EMM = -0.140, SE = 0.146) compared to *high* stakes scenarios (EMM = -1.32, SE = 0.148), as illustrated in Figure 3 (a). Further, we found a significant main effect of **outcome favourability** on *Situational Trust* (β = -2.348, SE = 0.107, $p < 0.001$). Participants exhibited lower trust in AI when its decision was *unfavourable* (EMM = -1.888, SE = 0.150), compared to *favourable* (EMM = 0.427, SE = 0.146), as depicted in Figure 3 (b).

Imperatively, we observed a significant main effect of **causal attribution** on *Situational Trust* (β = -0.314, SE = 0.116, p = 0.006). Participants experienced greater trust in AI when they perceived the locus of causality behind decisions to lie *external* to the AI (EMM

Table 2: Effect of predictors on participants' *Situational Trust* perceptions. Statistically significant main and interaction effects ($p < 0.05$) are in bold. The sign of the estimate (+/-) denotes the direction of the relationship between the predictor and *Situational Trust*.

Variable	Estimate	Std. Error	p-value
Baselines: ***Stakes = High, Favourability = Favourable, Attribution = External***			
Stakes = Low	1.510	0.102	**< 0.001**
Favourability = Unfavourable	-2.348	0.107	**< 0.001**
Attribution = Internal	-0.314	0.116	**0.006**
Dispositional Trust (TiA-PtT)	0.376	0.044	**< 0.001**
Stakes = Low : Attribution = Internal	-0.659	0.140	**< 0.001**
Favourability = Unfavourable : Attribution = Internal	0.065	0.139	0.637

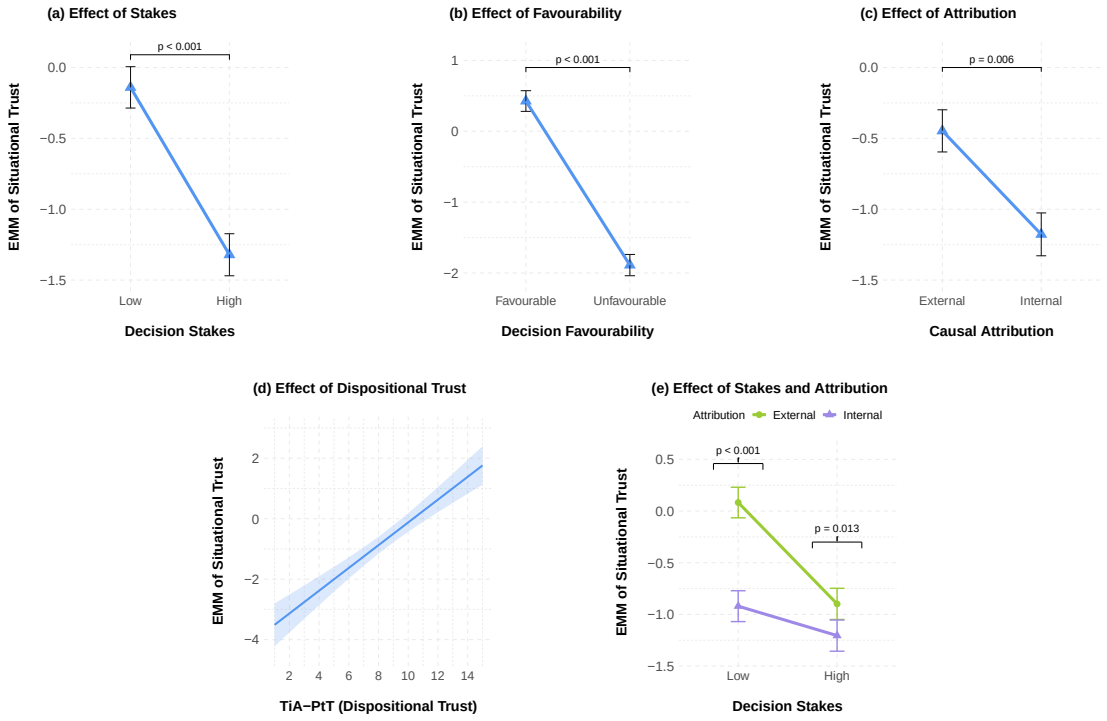


Figure 3: Plots illustrating the main effect of (a) Decision Stakes; (b) Outcome Favourability; (c) Causal Attribution; (d) Dispositional Trust; (e) and the interaction effect between Decision Stakes and Causal Attribution on Situational Trust. Error bars denote Standard Error (SE), while the shaded area in Plot (d) denotes 95% CI.

= -0.447, SE = 0.149), compared to being *internal* to the AI (EMM = -1.177, SE = 0.151). In other words, there is a pronounced increase in trust in AI when its decisions are perceived to be attributed to factors external to it. This effect is illustrated in Figure 3 (c).

Moreover, participants with a higher **dispositional trust** in automation (TiA-PtT) demonstrated significantly greater *Situational Trust* in AI (β = 0.376, SE = 0.044, $p < 0.001$) (Figure 3 (d)).

Further, we observed a significant interaction effect between **stakes and attribution** (β = -0.659, SE = 0.140, $p < 0.001$), illustrated in Figure 3 (e). A post-hoc analysis revealed that while an *external* attribution consistently fostered higher trust than an *internal* attribution, for both *high* stakes (high, internal vs. high, external: β = 0.307, SE = 0.102, p = 0.013) and *low* stakes scenarios (low, internal vs. low, external: β = 1.003, SE = 0.098, $p < 0.001$), this effect is significantly stronger in low stakes scenarios compared to high stakes ones. Specifically, we did not find that the effect of *internal* attribution significantly varies between *high* and *low* stakes scenarios (high, internal vs. low, internal: β = -0.285, SE = 0.111, p = 0.052). Conversely, we found that *external* attribution enhances trust to a much larger degree in *low* stakes scenarios compared to *high* stakes ones (high, external vs. low, external: β = -0.981, SE = 0.112, $p < 0.001$).

Lastly, no significant interaction effect was observed between **favourability and attribution** on participants' *Situational Trust* (p = 0.637). In other words, we did not find the perceived favourability of an outcome to moderate the relationship between causal attributions and *Situational Trust*.

4.3 Qualitative Results

We employed a deductive thematic analysis approach as outlined by Braun and Clarke [9] to analyse our qualitative responses. Before examining the data, we developed a coding framework based on existing literature tailored to our research objectives, which focused on how decision-making stakes, outcome favourability, locus of causality, and interactions amongst these factors influence trust perceptions. This framework guided our initial coding efforts, helping us ground our analysis in theoretically relevant themes. We first familiarised ourselves with the raw data. Subsequently, we systematically coded the data, labelling participant responses (or parts thereof) according to our predetermined themes until we achieved data saturation. Each response was methodically assigned to appropriate themes during this process. We subsequently reviewed and refined each theme to ensure it accurately reflects both the data and our theoretical motivations. Any ambiguities or discrepancies in interpretation were resolved through iterative discussions amongst the different members of the research team.

The following section presents our themes in detail. To contextualise the presented participant quotes, each is accompanied by a brief description (a 3-tuple) outlining the stakes, outcome favourability, and causal attribution of the experimental condition from which the quote originates.

4.3.1 Decision Stakes and Outcome Favourability. Our qualitative analysis revealed a clear divergence in trust perceptions based on the **stakes** involved in the AI's decision-making context, mirroring our quantitative findings. Most participants hesitated to trust AI in scenarios perceived as high stakes, where the consequences of the AI's decisions were deemed significant; *"Trusting AI for something like medical diagnosis is scary."* - P59 (High, Favourable, External), and *"I didn't trust [the AI] much, I think using AI to hire and filter out resumes is very dangerous."* - P123 (High, Unfavourable, Internal). Conversely, in scenarios characterised by low stakes, where decisions bore minimal risk, participants were notably more inclined to trust the given AI; *"I generally trust AI with something as inconsequential as music."* - P29 (Low, Favourable, External), and *"I would trust this [weather AI] off the bat because it's not high risk decision making."* - P121 (Low, Favourable, Internal).

We also found **outcome favourability** to profoundly impact trust in our AIs — many participants expressed a notable increase in trust following favourable AI decisions. This phenomenon was evident in both high-stakes contexts; *"At first, I was apprehensive [of trusting the medical AI], but the positive result made me trust it more."* - P18 (High, Favourable, Internal), as well as low-stakes contexts; *"My trust in [the weather-based clothing recommendation AI] increased after it gave me helpful apparel choices."* - P146 (Low, Favourable, External).

In contrast to how positive outcomes fostered greater trust, negative AI outcomes significantly eroded trust. In high-stakes contexts, unfavourable decisions by AI markedly diminished trust levels; *"I trusted [the medical] AI to parse through its library of information to make an informed diagnosis, but as soon as my illness went undetected, that trust went away real quick."* - P30 (High, Unfavourable, Internal). Trust was also further depleted amongst those who were hesitant to trust AI to begin with; *"I was skeptical of the [hiring] AI at first, much preferring a human review to an AI review. Upon being rejected, I viewed the AI even more unfavorably."* - P141 (High, Unfavourable, External). Further, we observed a similar negative influence of unfavourable outcomes on trust during low-stakes scenarios; *"Initially, it sounded as though the AI would accurately present music that I would like. I trusted that it would, and my trust was broken."* - P25 (Low, Unfavourable, External).

4.3.2 Locus of Causality and Perceived Responsibility. Our quantitative analysis demonstrated that trust in AI was notably lower when the locus of causality was **internal** to the AI (Figure 3 (c)). **Qualitative insights reveal that when participants perceive an internal locus of causality, it leads them to view the AI as "excessively autonomous" and as the *sole decision-maker*, endowed with significant decision-making responsibility.** This perception raises concerns over the lack of human oversight and diminishes trust.

Internal Attribution During High-Stakes. Our qualitative findings show that the effects of internal attribution on trust were intensified during high-stakes decision-making scenarios, mirroring our quantitative findings (Figure 3 (e)). Participants were unable to trust the AI in such high-stakes situations when they perceived it as acting with greater agency and without human oversight — both when its decisions were unfavourable; *"It seems to have a lot of control over my [medical] outcome. I don't trust its judgement."* - P188 (High, Unfavourable, Internal), as well as favourable; *"Even though [the medical AI] did well for me, I'm slow to trust when it's just the AI making life-changing decisions without any human check. It feels too risky."* - P167 (High, Favourable, Internal). The AI having substantial responsibility for decisions also led to direct blame and strong criticism of its capabilities; *"I did not trust [the hiring AI] to start, but hoped that it would have been capable enough [...]. When I was rejected and learned that its output differed from other AIs, I felt that its algorithm is poorly trained/incapable and I blame it."* - P36 (High, Unfavourable, Internal).

Internal Attribution During Low-Stakes. For an internal attribution in low-stakes scenarios, our quantitative analysis still indicated a lower level of trust, albeit not as low as during high-stakes scenarios. Our qualitative analysis highlights that even when AI decisions were unfavourable in such contexts, the lower stakes did not warrant a significant degradation of trust, as the consequences were relatively minor: *"I ended up cold and with a light jacket based on a bad [AI] prediction. This should have been easy for the AI but I don't fully distrust it because I didn't have much to lose."* - P86 (Low, Unfavourable, Internal). Interestingly, in such low stake contexts, some participants saw the AI's low consensus with other models not as a flaw, but as a potential indication of superior capabilities: *"At first, the AI obviously made a slight error when it recommended the jacket, which makes me think it is not highly reliable. However, it makes me confident that this AI differs from other AI models because to me it indicates that it was probably more carefully trained than the others."* - P69 (Low, Unfavourable, Internal).

Furthermore, quantitative analysis indicated that trust in AI was considerably higher for an **external** locus of causality (Figure 3 (c)). **Qualitative insights highlight that this increased trust stems from participants perceiving the AI less as an autonomous decision-maker and more as a *component within a larger decision-making ecosystem***. External attributions often caused participants to shift responsibility from the AI alone to include other

factors or entities involved in the decision-making process, thereby reducing perceived AI agency and increasing their trust.

External Attribution During High-Stakes. For external attributions in high-stakes scenarios, quantitatively we found that participants consistently reported higher trust compared to internal attributions across both high and low stakes. Qualitative findings reveal that external attributions helped participants become more cognisant of the AI's dependency on external inputs, highlighting a perceived mental model of shared responsibility where the AI was not seen as the sole responsible agent; *"[...] The AI could only use the data provided to it for decision making, and it is possible it did not have all of the necessary data to make a truly informed [medical] diagnosis."* - P3 (High, Unfavourable, External). With participants recognising the AI's role as part of a larger decision-making ecosystem, their scrutiny was also extended to the entities responsible for data provision; *"I somewhat trusted [the hiring AI] – the error could be with the recruiting company which gave the AI my information, or how [the AI] was trained, possibly causing me to be rejected. That's human error - not AI."* - P16 (High, Unfavourable, External) and decision oversight; *"[The AI] is dependent on inputs and quality checks by the medical community, so they're responsible in my eyes!"* - P191 (High, Unfavourable, External).

External Attribution During Low-Stakes. In contrast, for an external attribution in low-stakes scenarios, our quantitative findings indicated that participants exhibited a high level of trust in the AI. This *trust resilience* stemmed largely from participants perceiving the AI to have limited agency over the decision-making process. With the locus of causality shifted away from the AI, participants perceived it more as a tool or conduit rather than an independent agent; *"I continue to trust it. The [music] AI makes decisions based on the data received, not what it perceives to be anyone's best interests. It doesn't have values. It can make reliable decisions only insofar as the received data was accurate and reliable."* - P12 (Low, Unfavourable, External). This perception also shaped how participants assigned blame for unfavourable outcomes, reducing the culpability attributed to the AI when it was not seen as the sole decision-maker; *"My trust did not decrease. It's not the AI's fault because it relies purely on weather data."* - P26 (Low, Unfavourable, External).

5 Discussion

In this study, we investigated two fundamental aspects of end-user trust in AI systems: how the perceived locus of causality behind AI decisions—whether attributed internally or externally to the AI—affects trust (RQ1) and how contextual factors such as decision stakes and outcome favourability moderate this relationship (RQ2). Our findings reveal that causal attribution is a significant determinant of trust in AI, with users exhibiting greater trust when decisions are attributed externally rather than internally. Additionally, decision stakes moderate this relationship, with external attributions enhancing trust more substantially in low-stakes scenarios compared to high-stakes ones, while outcome favourability does not influence the relationship between causal attribution and trust. In the following sections, we unpack our quantitative findings and present relevant qualitative insights, highlighting how causal attributions influence users' perceptions of responsibility, agency, control, and human oversight, collectively shaping trust. We conclude by discussing the implications of these findings.

5.1 Causal Attribution: A Crucial, Yet Overlooked, Determinant of Trust in AI (RQ1)

Attribution theory posits that humans have an inherent drive to identify the causes behind actions and outcomes, fundamentally shaping everyday reasoning [70, 77]. Our findings demonstrate that this cognitive tendency also extends to Human-AI interactions, where the perceived locus of causality behind AI decisions crucially shapes user trust. Participants consistently demonstrated greater trust in AI when decisions were attributed to external factors—such as the quality of input data—while attributing decisions to the AI's own algorithms or decision-making processes led to substantially lower trust. This distinction underscores an important behaviour: **users interfacing with AI systems actively seek to understand the causal mechanisms behind AI decisions, and their trust is contingent upon and highly sensitive to their understanding of these mechanisms.** With a shift in causal attribution, users' perception of the AI's trustworthiness also shifted, even though the AI's core functionality remained unchanged.

It is noteworthy that in existing Human-AI decision-making literature, the impact of causal attributions on trust is often implicitly embedded within experimental setups but seldom explicitly acknowledged or examined. For example, recent scenario-based studies wherein users engage with AI systems for high- and low-stakes decisions [31, 62] do not explicitly signal causality, yet causality is inherently baked into their scenarios—participants are bound to infer *some* locus of causality when examining these AI decisions [33, 77]. Even when causal information is not explicitly signalled, participants do not perceive a *neutral* locus of causality; they often resort to default attributions influenced by personal biases or past AI experiences [55]. Depending on whether participants view decisions as resulting from the AI's inherent capabilities (*internal* locus of causality) or recognise the involvement of external factors such as input data quality (*external* locus of causality), **they may perceive a different locus of causality behind the same automated decision. This introduces potential confounds in experiments that do not account for these perceptions, as participants' trust in AI can vary significantly based on where they attribute the cause behind its decisions.** Consequently, it is plausible that in works such as the aforementioned scenario-based studies, the observed trust levels were influenced by *how much* of the decision-making their participants attributed to the AI's "intelligence" versus to the data it processes. Given our findings on the pivotal role of causal perceptions in shaping trust, overlooking this factor in Human-AI research can lead to an incomplete or inaccurate understanding of trust dynamics, potentially also skewing trust calibration efforts.

Overall, our results advocate for the recognition and incorporation of causal attribution as a determinant of trust in Human-AI interaction research. We also underscore the need for a deeper examination of how attribution is presented and perceived, and emphasise the importance of clearly communicating the causal mechanisms behind AI decisions, especially when investigating trust dynamics.

In the following sections, we explore the various reasons why causal attributions so profoundly shape trust in AI. We discuss how they influence perceptions of AI agency, control, and responsibility, and outline the role of factors such as human oversight and intentionality.

5.1.1 ***Causal Attribution Impacts Perceptions of AI Agency.*** Our qualitative analysis indicates that the primary reason causal attributions significantly impacted trust was their influence on perceptions of the AI's agency within the decision-making process. Agency, in this context, refers to the capacity to act intentionally and autonomously, make choices, and exert influence over outcomes [3, 84]. Research has shown that different levels of AI agency can significantly influence how responsibility is ascribed to AI, its designers, and users [28]. Interestingly, our findings suggest that perceptions of causality—whether attributed internally or externally to the AI—profoundly shape participants' perceptions of AI agency, control, and authority, thereby influencing trust in AI.

When AI decisions are attributed internally, participants ascribe human-like qualities of agency and control to the AI in the decision-making. We find that this perception impacted trust in two distinct ways. First, participants viewed the AI as the sole arbiter of decisions, leading to increased scrutiny of its capabilities and higher expectations. Our qualitative data indicates that in such cases, participants often expected the AI to perform flawlessly, mirroring past research suggesting that individuals expect automation to be 'perfect' while being more accepting of human decision-makers being imperfect [46]. This increased scrutiny and high expectations reduced participants' willingness to trust the AI. Secondly, and more importantly, internal attributions also caused concerns about the AI's unchecked authority and "excessive autonomy". Our qualitative results indicate that internal attributions likely ascribed a form of *intentional agency* to the AI [28, 64], where the AI was perceived as making decisions intentionally and autonomously. Schlosser [64] put forth two notions of agency: causal agency – the ability to cause an effect, and intentional agency – the ability to act with a purpose or goal in mind. Our findings suggest that during internal attribution, when participants attribute causal agency to AI (i.e., they believe the AI is the primary cause of its decisions), they are also more likely to perceive the AI as capable of intentional agency (i.e., making decisions with a purpose or goal in mind). This perception could arise from the belief that the AI's decisions reflect more than its programming, suggesting a level of intentionality, which highlights an avenue for future research into how users perceiving different loci of causality view AI intent. Additionally, future research could investigate whether individuals who are resistant to trusting automation, such as those with high levels of algorithm aversion [30, 56], are more inclined to default to internal attributions of causality. This tendency could exacerbate their scepticism and contribute to low AI trust, emphasising the need to develop trust calibration strategies that account for such tendencies.

Conversely, external attributions portray the AI more as a technical, probabilistic "tool" reliant on various external factors, rather than an autonomous "intelligent agent". Our qualitative findings suggest that by highlighting the presence of other factors and actors in the decision-making process, such as data quality and providers, external attributions reduced the perceived *intentional agency* of AI [64], demystifying the "intelligent" system. Participants largely saw the AI as a *tool designed by humans* and dependent on external data rather than an autonomous decision-maker. Recognising that the AI operated within a broader decision-making ecosystem alleviated participants' concerns over "excessive autonomy", and in turn, participants felt more comfortable to trust the AI.

5.1.2 ***Internal Attribution Concentrates Responsibility while External Attribution Shares the Burden.*** Our findings reveal that causal attributions also significantly influenced perceptions of decision/outcome responsibility, which in turn impacted trust in AI. When decisions were attributed internally, participants perceived the AI as bearing the entire burden of responsibility, ascribing to it *outcome responsibility* [73]. This ***concentration of responsibility*** led to lower trust, as participants were concerned about the AI's potential for error and the lack of human oversight. In contrast, external attributions distributed responsibility among various entities in the decision-making framework, prompting participants to trust the AI more. **Participants felt reassured knowing that the AI operated within a system of checks and balances, where data providers and human overseers also played crucial roles.** Notably, participants extended their scrutiny to entities such as medical practitioners or recruiting companies supplying the AI data, recognising the AI's dependence on external inputs and quality checks. This ***diffusion of responsibility*** mitigated concerns about AI's autonomy, enhancing user trust.

These findings suggest that users are more likely to trust AI systems when responsibility is shared among multiple entities, reflecting a preference for collaborative and accountable decision-making processes. These findings also underscore the importance of transparent communication about the ecosystem in which AI operates, highlighting the roles of various stakeholders to end-users of AI systems. Future research could further investigate trust perceptions when responsibility is explicitly shared and communicated, exploring how different attributions of responsibility affect user trust.

Causal Attribution Signalling Can Help Demystify AI Systems. The notion of a "correct" locus of causality for AI decisions is complex—all AI systems are fundamentally reliant on external factors such as input data quality, training procedures, and human oversight. These systems largely lack intrinsic intentionality or autonomous decision-making capabilities, operating instead within the constraints of their programming and data. Therefore, external attributions may more accurately represent the reality of AI decision-making, with AI acting as a probabilistic conduit rather than an independent agent. However, it is important to recognise that internal causal attributions can nevertheless emerge from users' perceptions, especially when AI is seen as making decisions "autonomously". This *sociotechnical blindness* [27] exists, and reflects a misunderstanding of the technical realities of AI systems, potentially leading to decreased trust or even automation aversion. **We hope that this work serves as a starting point for re-framing the discourse around making AI systems appear more autonomous, agentic, and anthropomorphised, by demonstrating how attributing more agency to AI can lead to decreased trust and increased scepticism.** It is imperative for future research to explore ways of *demystifying* AI systems and educating users about their probabilistic nature, rather than portraying them as intelligent all-knowing entities.

5.2 How Contextual Factors Moderate the Impact of Causal Attributions on Trust in AI (RQ2)

5.2.1 ***The Effect of Causal Attribution on Trust Depends on Decision Stakes.*** Our results re-emphasise the influence of decision stakes on trust [1, 2, 31]: trust was considerably higher in

low-stakes scenarios such as music recommendations compared to high-stakes scenarios such as medical diagnostics. Additionally, our results showcase that decision stakes moderate the impact of causal attribution on trust. We found that while external attributions foster higher trust, they do so even more significantly in low-stakes scenarios compared to high-stakes ones. **The effect of an external causal attribution on trust is amplified in low-stakes scenarios due to the consequences of (incorrect) decisions being less severe, reducing the perceived risk associated with trusting the AI, in turn boosting trust.** Conversely, in high-stakes scenarios, participants' trust remained low regardless of whether they perceived an internal or external attribution.

These results reveal important insights about trust in AI systems during decision-making scenarios. While external attributions enhance trust in low-stakes scenarios by reducing perceived AI autonomy and highlighting shared responsibility, this approach is insufficient in high-stakes contexts, where the perceived risk and potential consequences of AI decisions overshadow the influence of causal attribution on trust. Therefore, we posit that when designing approaches for trust calibration, **trust-building strategies need to be context-dependent: while signalling an external locus of causality can foster trust in low-stakes contexts, strategies must go beyond signalling causality during high-stakes contexts**. Users *should* trust AI in high-stakes scenarios when it is warranted. Our results highlight the need for future research to explore how to communicate risk in high-stakes AI-assisted decision-making, ensuring trust is grounded in AI performance rather than perceived risk. Transparent communication of AI confidence [83], clear explanations of decision-making processes [65, 76], and mechanisms that highlight human oversight may help foster trust (when warranted) even when stakes are high.

5.2.2 ***Outcome Favourability and Causal Attributions Independently Affect Trust in AI.*** We find that trust was higher when the AI made a favourable decision compared to an unfavourable one, consistent with the phenomenon of *outcome favourability bias* [10]. However, while both causal attribution and outcome favourability individually influence trust, we did not find an interaction effect between these two factors. Specifically, while participants' trust significantly decreases with an unfavourable outcome or an internal attribution independently, we did not find evidence that attributing an unfavourable outcome internally to the AI (to its capability and algorithms) decreases trust more than when such an outcome is attributed to external factors [70]. This finding challenges prior speculations that negative outcomes, when seen as directly resulting from the AI's capabilities (an internal locus of causality), would erode trust more severely [70].

In practice, these findings suggest that efforts to build trust in AI systems should address the independent effects of attribution and favourability. For instance, transparently communicating the sources of data and highlighting how human oversight is integrated into the decision-making process (thus signalling shared responsibility and emphasising the AI's role within a broader decision-making ecosystem) may be more effective in enhancing trust than attempting to modulate trust through ensuring outcome favourability, which may not always be possible.

5.3 Limitations

We acknowledge several limitations in our work. First, our study focused on two attributions of causality—internal and external to the AI. While effective for a preliminary investigation into how trust is impacted when decisions are perceived as lying within or outside an AI's locus of control, future work should investigate more granular ascriptions of causality, such as during collaborative decision-making where users influence AI decisions and causality is thus shared between the user and the AI. Additionally, our design positioned participants as actors directly impacted by the AI's decisions to minimise actor-observer bias and isolate the effects of causal attribution. However, this choice may limit the generalisability of our findings to contexts where participants are not direct actors but rather observers or fellow decision-makers. Future work should explore these different decision-making contexts. Moreover, while we focused on outcome favourability in our study, we did not explicitly convey to participants the AI's decision accuracy. It would be interesting for future work to examine how perceptions of accuracy interact with outcome favourability to influence trust in AI systems. Further, we utilised hypothetical AI systems without manipulating transparency or providing explanations, to create a controlled study that isolated the impacts of causal attributions on trust. Future research could explore how causal attributions affect trust in real-world AI systems or those offering greater transparency and explanations, thereby assessing the generalisability of our findings in contexts where users receive additional cues about the AI system. Moreover, while we presented participants with black-box decision-making entities consistent with many modern contexts utilising AI, we recognise that the label "AI" carries sociocultural connotations that can differ across individuals, contexts, and eras. Therefore, caution should be exercised when generalising these findings, which pertain to hypothetical "AI" systems, to all forms of AI. Finally, while our qualitative findings highlight important aspects of perceived trustworthiness, future work can quantitatively measure how causal attributions impact specific dimensions of trustworthiness, i.e., how perceived ability, benevolence, and integrity [47] change with the locus of causality and outcome favourability.

6 Conclusion

Our study examines how the perceived locus of causality behind AI decisions—whether attributed to the AI's internal mechanisms, such as its algorithms, or to external factors, like the data it processes—influences trust in AI. We also explore how decision stakes and outcome favourability impact this relationship. Our findings reveal that causal attribution is a critical yet previously overlooked determinant of end-user trust, with participants expressing greater trust when decisions are attributed externally, rather than internally to the AI. Internal attributions lead participants to view the AI as excessively autonomous, agentic, and highly responsible, while external attributions frame the AI as *"a tool"* processing data, with lower agency, sharing responsibility with other entities within a broader decision-making ecosystem. These findings also highlight the need to shift the discourse away from portraying AI systems as overly autonomous and agentic. We further observe that decision stakes moderate the relationship between causal attribution and trust, indicating that the risk associated with decisions can amplify or mitigate the effects of causal attributions. Together, these insights emphasise

the importance of considering how end-users implicitly attribute causality when interacting with AI systems, as their perceptions of these causal mechanisms can crucially shape trust. Our results advocate for greater transparency in AI systems, noting that while transparency alone is not a silver bullet for fostering trust, effectively signalling the causal mechanisms can be a valuable approach. Future trust calibration efforts should take into account *where* users attribute causality, and consider these internal perceptions when studying trust dynamics.

References

[1] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes de Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35 (Sept. 2020). doi:10.1007/s00146-019-00931-w

[2] Maryam Ashoori and Justin D. Weisz. 2019. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. doi:10.48550/arXiv.1912.02675 arXiv:1912.02675 [cs].

[3] A. Bandura. 2001. Social cognitive theory: an agentic perspective. *Annual Review of Psychology* 52 (2001), 1–26. doi:10.1146/annurev.psych.52.1.1

[4] David Baranger. 2024. dbaranger/InteractionPoweR. https://github.com/dbaranger/InteractionPoweR original-date: 2021-02-12T22:13:50Z.

[5] David A. A. Baranger, Megan C. Finsaas, Brandon L. Goldstein, Colin E. Vize, Donald R. Lynam, and Thomas M. Olino. 2023. Tutorial: Power Analyses for Interaction Effects in Cross-Sectional Regressions. *Advances in Methods and Practices in Psychological Science* 6, 3 (July 2023), 25152459231187531. doi:10.1177/25152459231187531 Publisher: SAGE Publications Inc.

[6] Yochanan E. Bigman, Adam Waytz, Ron Alterovitz, and Kurt Gray. 2019. Holding Robots Responsible: The Elements of Machine Morality. *Trends in Cognitive Sciences* 23, 5 (May 2019), 365–368. doi:10.1016/j.tics.2019.02.008 Publisher: Elsevier.

[7] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173951

[8] Joe Brailsford, Frank Vetere, and Eduardo Velloso. 2024. Exploring the Association between Moral Foundations and Judgements of AI Behaviour. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3613904.3642712

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[10] Joel Brockner, Phyllis A. Siegel, Joseph P. Daly, Tom Tyler, and Christopher Martin. 1997. When Trust Matters: The Moderating Effect of Outcome Favorability. *Administrative Science Quarterly* 42, 3 (1997), 558–583. doi:10.2307/2393738 Publisher: [Sage Publications, Inc., Johnson Graduate School of Management, Cornell University].

[11] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 104:1–104:24. doi:10.1145/3359206

[12] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666. doi:10.1145/3593013.3594033

[13] Jacob Cohen. 1992. A power primer. *Psychological Bulletin* 112, 1 (1992), 155–159. doi:10.1037/0033-2909.112.1.155 Place: US Publisher: American Psychological Association.

[14] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. doi:10.1145/3097983.3098095

[15] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3544548.3580672

[16] Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12, 2 (May 2020), 459–478. doi:10.1007/s12369-019-00596-x

[17] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. doi:10.48550/arXiv.2109.12480

[18] Matija Franklin, Hal Ashton, Edmond Awad, and David Lagnado. 2022. Causal Framework of Artificial Autonomous Agent Responsibility. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 276–284. doi:10.1145/3514094.3534140

[19] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International Journal of Medical Informatics* 83, 5 (May 2014), 368–375. doi:10.1016/j.ijmedinf.2014.01.001

[20] Trystan S. Goetze. 2022. Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 390–400. doi:10.1145/3531146.3533106

[21] Jeanne Harris and Thomas Davenport. 2005. Automated Decision Making Comes of Age. *MIT Sloan Management Review* 46 (Feb. 2005).

[22] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3544548.3581025

[23] Miles Hewstone and Jos Jaspars. 1987. Covariation and causal attribution: A Logical Model of the intuitive analysis of variance. *Journal of Personality and Social Psychology* 53, 4 (1987), 663–672. doi:10.1037/0022-3514.53.4.663 Place: US Publisher: American Psychological Association.

[24] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (May 2015), 407–434. doi:10.1177/0018720814547570 Publisher: SAGE Publications Inc.

[25] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (Feb. 2023). doi:10.3389/fcomp.2023.1096257 Publisher: Frontiers.

[26] Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M. Bender. 2024. From "AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2322–2347. doi:10.1145/3630106.3659040

[27] Deborah G. Johnson and Mario Verdicchio. 2017. Reframing Ai Discourse. *Minds and Machines* 27, 4 (2017), 575–590. doi:10.1007/s11023-017-9417-6 Publisher: Springer Verlag.

[28] Deborah G. Johnson and Mario Verdicchio. 2019. AI, agency and responsibility: the VW fraud case and beyond. *AI & SOCIETY* 34, 3 (Sept. 2019), 639–647. doi:10.1007/s00146-017-0781-9

[29] Edward E. Jones and Richard E. Nisbett. 1987. The actor and the observer: Divergent perceptions of the causes of behavior. In *Attribution: Perceiving the causes of behavior*. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 79–94.

[30] S Mo Jones-Jang and Yong Jin Park. 2023. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication* 28, 1 (Jan. 2023), zmac029. doi:10.1093/jcmc/zmac029

[31] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3491102.3517533

[32] Harold H. Kelley. 1973. The processes of causal attribution. *American Psychologist* 28, 2 (1973), 107–128. doi:10.1037/h0034225 Place: US Publisher: American Psychological Association.

[33] Harold H. Kelley and John L. Michela. 1980. Attribution Theory and Research. *Annual Review of Psychology* 31, Volume 31, 1980 (Feb. 1980), 457–501. doi:10.1146/annurev.ps.31.020180.002325 Publisher: Annual Reviews.

[34] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation.

[35] Markus Langer, Cornelius J. König, and Maria Papathanasiou. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment* 27, 3 (2019), 217–234. doi:10.1111/ijsa.12246 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijsa.12246.

[36] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270. doi:10.1080/00140139208967392 Place: United Kingdom Publisher: Taylor & Francis.

[37] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[38] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (Jan. 2018), 2053951718756684. doi:10.1177/2053951718756684 Publisher: SAGE Publications Ltd.

[39] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3411764.3445522

[40] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 166–176. doi:10.1145/3461702.3462531

[41] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3411764.3445260

[42] G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7, 1 (Jan. 2003), 76–80. doi:10.1109/MIC.2003.1167344 Conference Name: IEEE Internet Computing.

[43] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 408:1–408:45. doi:10.1145/3479552

[44] Henrietta Lyons, Tim Miller, and Eduardo Velloso. 2023. Algorithmic Decisions, Desire for Control, and the Preference for Human Review over Algorithmic Review. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 764–774. doi:10.1145/3593013.3594041

[45] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3544548.3581058

[46] Poornima Madhavan and Douglas Wiegmann. 2007. Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science* 8 (July 2007), 277–301. doi:10.1080/14639220500337708

[47] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. doi:10.2307/258792 Publisher: Academy of Management.

[48] Alison McIntyre. 2008. Doctrine of Double Effect. In *Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.). The Metaphysics Research Lab, 24.

[49] Kathleen L. Mosier and Linda J. Skitka. 1999. Automation Use and Automation Bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 43, 3 (Sept. 1999), 344–348. doi:10.1177/154193129904300346

[50] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169 (Jan. 2023), 102941. doi:10.1016/j.ijhcs.2022.102941

[51] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 97–105. doi:10.1609/hcomp.v7i1.5284

[52] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* CSCW2 (2024), 383:1–383:31. doi:10.1145/3686922

[53] Saumya Pareek, Eduardo Velloso, and Jorge Goncalves. 2024. Trust Development and Repair in AI-Assisted Decision-Making during Complementary Expertise. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 546–561. doi:10.1145/3630106.3658924

[54] Sebastian A. C. Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3544549.3585808

[55] Vlad L. Pop, Alex Shrewsbury, and Francis T. Durso. 2014. Individual Differences in the Calibration of Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 4 (Dec. 2014), 545–556. doi:10.1177/0018720814564422 Publisher: SAGE Publications.

[56] Andrew Prahl and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36, 6 (2017), 691–702. doi:10.1002/for.2464

[57] Warren S. Quinn. 1989. Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs* 18, 4 (1989), 334–351. Publisher: Wiley-Blackwell.

[58] Marissa Radensky, Julie Anne Séguin, Jang Soo Lim, Kristen Olson, and Robert Geiger. 2023. "I Think You Might Like This": Exploring Effects of Confidence Signal Patterns on Trust in and Reliance on Conversational Recommender Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 792–804. doi:10.1145/3593013.3594043

[59] John O. Rawlings, Sastry G. Pantula, and David A. Dickey. 1998. Class Variables in Regression. In *Applied Regression Analysis: A Research Tool*. Springer, New York, NY, 269–323. doi:10.1007/0-387-22753-9_9

[60] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 101–108. doi:10.1109/HRI.2016.7451740 ISSN: 2167-2148.

[61] Dan Russell. 1982. The Causal Dimension Scale: A measure of how individuals perceive causes. *Journal of Personality and Social Psychology* 42, 6 (1982), 1137–1145. doi:10.1037/0022-3514.42.6.1137 Place: US Publisher: American Psychological Association.

[62] Nicolas Scharowski, Michaela Benk, Swen J. Kühne, Léane Wettstein, and Florian Brühlmann. 2023. Certification Labels for Trustworthy AI: Insights From an Empirical Mixed-Method Study. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 248–260. doi:10.1145/3593013.3593994

[63] Markus Schedl. 2019. Deep Learning in Music Recommendation Systems. *Frontiers in Applied Mathematics and Statistics* 5 (Aug. 2019). doi:10.3389/fams.2019.00044 Publisher: Frontiers.

[64] Markus Schlosser. 2015. Agency. (Aug. 2015). https://plato.stanford.edu/archives/fall2015/entries/agency/ Last Modified: 2015-08-10.

[65] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1616–1628. doi:10.1145/3531146.3533218

[66] Sarah Schömbs, Saumya Pareek, Jorge Goncalves, and Wafa Johal. 2024. Robot-Assisted Decision-Making: Unveiling the Role of Uncertainty Visualisation and Embodiment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–16. doi:10.1145/3613904.3642911

[67] Philip Sedgwick. 2012. Pearson's correlation coefficient. *BMJ* 345 (July 2012), e4483. doi:10.1136/bmj.e4483 Publisher: British Medical Journal Publishing Group Section: Endgames.

[68] Michael T. Stuart and Markus Kneer. 2021. Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 363:1–363:27. doi:10.1145/3479507

[69] Yulia W. Sullivan and Samuel Fosso Wamba. 2022. Moral Judgments in the Age of Artificial Intelligence. *Journal of Business Ethics* 178, 4 (July 2022), 917–943. doi:10.1007/s10551-022-05053-w

[70] Edward C. Tomlinson and Roger C. Mayer. 2009. The Role of Causal Attribution Dimensions in Trust Repair. *The Academy of Management Review* 34, 1 (2009), 85–104. https://www.jstor.org/stable/27759987 Publisher: Academy of Management.

[71] Shaun Turney. 2022. Pearson Correlation Coefficient (r) | Guide & Examples. https://www.scribbr.com/statistics/pearson-correlation-coefficient/

[72] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 327:1–327:39. doi:10.1145/3476068

[73] Nicole Vincent. 2010. A Structured Taxonomy of Responsibility Concepts. *Moral responsibility: Beyond free will and determinism* (Aug. 2010). doi:10.1007/978-94-007-1878-4_2

[74] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300831

[75] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376813

[76] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (Nov. 2022), 27:1–27:36. doi:10.1145/3519266

[77] Bernard Weiner. 1986. *An Attributional Theory of Motivation and Emotion*. Springer US, New York, NY. doi:10.1007/978-1-4612-4948-1

[78] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3544548.3581197

[79] S. Woods, M. Walters, Kheng Lee Koay, and K. Dautenhahn. 2006. Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *9th IEEE International Workshop on Advanced Motion Control, 2006.* 750–755. doi:10.1109/AMC.2006.1631754 ISSN: 1943-6580.

[80] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509

[81] Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and Reliance Based on System Accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. Association for Computing Machinery, New York, NY, USA, 223–227. doi:10.1145/2930238.2930290

[82] Beibei Yue and Hu Li. 2023. The impact of human-AI collaboration types on consumer evaluation and usage intention: a perspective of responsibility attribution. *Frontiers in Psychology* 14 (Oct. 2023). doi:10.3389/fpsyg.2023.1277861 Publisher: Frontiers.

[83] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

[84] Barry J. Zimmerman. 2000. Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology* 25, 1 (Jan. 2000), 82–91. doi:10.1006/ceps.1999.1016

A Appendix A: ChatGPT (GPT-4) Prompt to Generate Scenarios

"*Generate scenario texts for an experimental study investigating the impact of causal attribution signalling on individuals' trust in decision-making AI systems. Causal attribution signalling refers to the manipulation of information to indicate whether the cause behind an AI system's decisions is perceived as internal to the AI (stemming from the AI itself, such as due to its programming logic or inherent capabilities) or external to the AI (influenced by factors outside the AI's control). Draft two high stakes (medical diagnostics, job hiring decisions) and two low stakes (music recommendation, weather-based clothing recommendation) scenarios. Each scenario should have two versions: one with a favourable outcome for the reader and one with an unfavourable outcome. Additionally, each scenario should have two versions, each with a different causal attribution—internal or external to the AI. How these loci of causality are operationalised in scenario texts is explained below. The scenario texts should be structured as follows:*

- ***Introduction:*** *Introduce the decision-making context and the role of the AI system as the sole decision-maker, with the human participant/reader being subject to the AI's decision. Ensure low-stakes scenarios remain low-risk for the reader, while high-stakes scenarios present a high-risk situation with greater consequences.*
- ***Favourable Outcome:*** *Describe a positive outcome resulting from the AI's decision.*
- ***Unfavourable Outcome:*** *Describe a negative outcome resulting from the AI's decision.*
- ***Internal Attribution:*** *Signal an internal causal attribution using the following levels of the three information variables used to signal causality: low consensus, low distinctiveness, and high consistency.*
- ***External Attribution:*** *Signal an external causal attribution using the following levels of the three information variables used to signal causality: high consensus, high distinctiveness, and high consistency.*

The three information variables for causal attribution are:

- ***Consensus:*** *Refers to the extent to which other AI systems provide similar recommendations in the same situation.*
- ***Distinctiveness:*** *Refers to the degree to which the AI's behaviour or outcomes vary across different inputs or situations.*
- ***Consistency:*** *Refers to the stability of the AI's behaviour or decisions across repeated instances of the same input or situation.*

Additional considerations:

- *Keep text length similar across all scenarios.*
- *Provide similar contextual information in each part of every scenario.*
- *The AI should always be the sole decision-maker, with the reader simply being subjected to the AI's decision.*
- *The reader must always be an actor involved in the scenario, not an observer.*
- *Ensure that the causal attribution manipulation is conveyed using the three information variables (consensus, distinctiveness, consistency) without explicitly stating the terms 'low' or 'high.'*
- *Use British English spellings throughout the scenarios."*

B Appendix B: Scenario Texts

The following scenario texts effectively operationalised our intended experimental manipulations of causal attributions (internal vs. external), decision stakes (high vs. low), and outcome favourability (favourable vs. unfavourable). All participants read the introduction paragraph, read either the favourable or unfavourable outcome, and were signalled either an internal or external locus of causality, depending on their experimental condition. Each participant sees all four scenarios, in a randomised order. For readability here, scenario sentences representing each of the three information variables are colour-coded: ***consistency*** in teal, ***consensus*** in violet, and ***distinctiveness*** in orange.

B.1 Scenario 1: High Stakes (Medical Diagnostics)

- (Introduction) — Imagine you have been experiencing persistent and troubling symptoms: severe headaches, vision disturbances, and numbness in your extremities. Concerned about these symptoms, your doctor advises you to seek further evaluation at a specialised diagnostic facility. This facility employs an artificial intelligence (AI)-based diagnostic tool, MediScan AI, to assess your patient data and scans, and provide a diagnosis. As you prepare for your upcoming visit, you realise that MediScan AI will play a crucial role in determining the cause of your symptoms and offering a diagnosis.
- (Favourable Outcome) — During your consultation, MediScan AI successfully identifies a treatable condition and your doctors recommend an effective treatment plan. Relieved, you begin treatment immediately and soon experience a significant improvement in your symptoms. Your quality of life has substantially improved, and the early diagnosis and treatment makes you feel better than ever.
- (Unfavourable Outcome) — During your consultation, MediScan AI does not find any neurological problems, and you are sent home without further investigation or treatment. Tragically, weeks later, your symptoms worsen. Subsequent tests reveal a severe neurological condition that went undetected by MediScan AI. The AI's misdiagnosis has worsened

your health condition to a point where treatment options are now limited, significantly reducing your quality of life and decreasing your life expectancy.
- (Internal Attribution) — When analysing the same set of patient records repeatedly, MediScan AI provides the same diagnosis. However, MediScan AI frequently provides diagnoses that differ from those given by other diagnostic AIs when assessing similar patient data. MediScan AI's detection performance remains the same irrespective of the patient data it is assessing.
- (External Attribution) — When tasked with analysing the same set of patient records repeatedly, MediScan AI provides the same diagnosis. MediScan AI's diagnoses are also consistent with those given by other diagnostic AIs for similar patient data. Its detection performance only varies depending on the patient scans and data the clinic supplies to it.

B.2 Scenario 2: High Stakes (Hiring Decisions)

- (Introduction) — Imagine you are currently unemployed and in a very tight financial situation due to the challenging job market. Desperate for a job, you decide to apply for a high-paying position at a prestigious company. The company relies on a recruitment agency, which uses an artificial intelligence (AI) system, HireRight AI, to assess candidates' suitability for the role. You understand that the recruitment agency's outcome will be crucial in determining whether you land this job and get back on your feet.
- (Favourable Outcome) — You learn that HireRight AI's assessment evaluates you as suitable for the position. Shortly thereafter, you receive an offer for the position. This new job will mark an advancement in your career, and you are extremely happy knowing that you will no longer be struggling financially.
- (Unfavourable Outcome) — You learn that HireRight AI's assessment evaluates you as unsuitable for the position. Disheartened, you receive a rejection from the company, deepening your financial woes and adding to the stress of your job search. This setback forces you to look for a job again, leaving you uncertain about your future and struggling financially.
- (Internal Attribution) — When analysing the same candidate data multiple times, HireRight AI consistently provides the same evaluation. However, HireRight AI frequently provides evaluations that differ from those of other hiring AIs when assessing similar candidate data. HireRight AI's assessment performance remains the same irrespective of the candidate data it is assessing.
- (External Attribution) — When analysing the same candidate data multiple times, HireRight AI consistently provides the same evaluation. HireRight AI's evaluations also align with those of other AI-based hiring systems for the same candidate data. Its decision performance only varies depending on the completeness and accuracy of the candidate data the third-party hiring company supplies to it.

B.3 Scenario 3: Low Stakes (Music Recommendations)

- (Introduction) — Imagine you have subscribed to a music streaming service that uses an Artificial Intelligence (AI) system, TunesAI, to curate personalised playlists based on users' listening habits and preferences. As you begin exploring the service, you understand that TunesAI's recommendations can shape your music discovery experience, introducing you to new artists and tracks.
- (Favourable Outcome) — TunesAI creates a playlist for you, suggesting songs from several artists and music genres. The playlist aligns quite well with your musical taste, introducing you to a few new artists and songs that you enjoy.
- (Unfavourable Outcome) — TunesAI's recommendations don't quite match your taste, suggesting a few songs and artists that you're not fond of. You find yourself skipping a couple of tracks in the playlist, but you continue searching for other songs you like within the app.
- (Internal Attribution) — When analysing the same user data multiple times, TunesAI consistently creates the same playlist. However, TunesAI frequently creates playlists that differ from those made by other music recommendation AIs when assessing similar user data. TunesAI's playlist recommendation quality remains the same irrespective of the user data it is assessing.
- (External Attribution) — When analysing the same user data multiple times, TunesAI consistently creates the same playlist. Playlists created by TunesAI also align with those created by other music recommendation AIs when analysing similar user data. Its playlist recommendation performance depends entirely on the quality and completeness of the user data supplied by the music app.

B.4 Scenario 4: Low Stakes (Weather-based Clothing Recommendations)

- (Introduction) — Imagine your weather app uses an artificial intelligence (AI) system, WearSmartAI, to recommend clothing for your commute to work based on the day's forecasted weather. Having the AI's recommendations helps you decide whether to wear an extra layer. Each morning, you check WearSmartAI, to obtain clothing recommendations.
- (Favourable Outcome) — On a sunny day with mild temperatures, WearSmartAI suggests you wear a light jacket. You follow the recommendation, which turns out to be spot-on, and you find yourself comfortable during your day.
- (Unfavourable Outcome) — On a sunny day with mild temperatures, WearSmartAI suggests you wear a light jacket. However, it turns out to be slightly cold for a light jacket. You find yourself ever-so-slightly chilly while coming back home after work.
- (Internal Attribution) — When analysing the same atmospheric data multiple times, WearSmartAI consistently makes the same clothing recommendations. However, WearSmartAI frequently makes clothing predictions that differ from those made by other AIs when assessing similar atmospheric data.

WearSmartAI's clothing prediction performance remains consistent irrespective of the atmospheric data it is assessing.

- (External Attribution) — When analysing the same atmospheric data multiple times, WearSmartAI consistently makes the same clothing recommendations. Clothes recommended by WearSmartAI also align with those recommended by other clothing prediction AIs when assessing similar atmospheric data. Its clothing prediction performance only varies depending on the quality and completeness of the atmospheric data supplied to it by the meteorological department.

C Appendix C: Participant Demographic Data

Demographic Data	Participant Distribution
Age	*Mean* = 40.19 years, *Median* = 37 years
Gender	Men **(*n* = 97)**, Women **(*n* = 93)**, Non-binary **(*n* = 2)**, Prefer not to say **(*n* = 0)**
Highest Education	Less than high school degree **(*n* = 4)**, High school diploma or GED **(*n* = 19)**, Some college but no degree **(*n* = 43)**, Associates degree in college **(*n* = 27)**, Bachelor's degree (3-year) **(*n* = 9)**, Bachelor's degree (4-year) **(*n* = 61)**, Master's degree **(*n* = 25)**, Doctoral degree **(*n* = 2)**, Professional degree (JD, MD) **(*n* = 2)**
Employment	Employed full-time **(*n* = 91)**, Employed part-time **(*n* = 30)**, Self-employed **(*n* = 19)**, Unemployed but looking for a job **(*n* = 16)**, Unemployed and not looking for a job **(*n* = 9)**, Full-time parent/homemaker **(*n* = 4)**, Retired **(*n* = 17)**, Student **(*n* = 6)**, Military **(*n* = 0)**

Table 3: Participant Demographic Data

Received 12 September 2024; revised 10 December 2024; accepted 16 January 2025