# Integrity-based Explanations for Fostering Appropriate Trust in AI Agents

Mehrotra, Siddharth; Centeio Jorge, Carolina; Jonker, Catholijn M.; Tielman, Myrthe L.

**DOI**
10.1145/3610578

**Publication date**
2024

**Document Version**
Final published version

**Published in**
ACM Transactions on Interactive Intelligent Systems

# Integrity-based Explanations for Fostering Appropriate Trust in AI Agents

SIDDHARTH MEHROTRA, Delft University of Technology, The Netherlands
CAROLINA CENTEIO JORGE, Delft University of Technology, The Netherlands
CATHOLIJN M. JONKER, Delft University of Technology & LIACS, Leiden University, The Netherlands
MYRTHE L. TIELMAN, Delft University of Technology, The Netherlands

Appropriate trust is an important component of the interaction between people and AI systems, in that "inappropriate" trust can cause disuse, misuse, or abuse of AI. To foster appropriate trust in AI, we need to understand how AI systems can elicit appropriate levels of trust from their users. Out of the aspects that influence trust, this article focuses on the effect of showing integrity. In particular, this article presents a study of how different integrity-based explanations made by an AI agent affect the appropriateness of trust of a human in that agent. To explore this, (1) we provide a formal definition to measure appropriate trust, (2) present a between-subject user study with 160 participants who collaborated with an AI agent in such a task. In the study, the AI agent assisted its human partner in estimating calories on a food plate by expressing its integrity through explanations focusing on either honesty, transparency, or fairness. Our results show that (a) an agent who displays its integrity by being explicit about potential biases in data or algorithms achieved appropriate trust more often compared to being honest about capability or transparent about the decision-making process, and (b) subjective trust builds up and recovers better with honesty-like integrity explanations. Our results contribute to the design of agent-based AI systems that guide humans to appropriately trust them, a formal method to measure appropriate trust, and how to support humans in calibrating their trust in AI.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Artificial intelligence**; **Intelligent agents**;

Additional Key Words and Phrases: Integrity, appropriate trust, trust, explanations, honesty, transparency, fairness, artificial agents, intelligent agents, HCI

**ACM Reference format:**
Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. Integrity-based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Trans. Interact. Intell. Syst.* 14, 1, Article 4 (January 2024), 36 pages.
https://doi.org/10.1145/3610578

This research was (partly) funded by the Hybrid Intelligence Center, a 10-year programme funded the Dutch Ministry of Education, Culture, and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022, and by EU H2020 ICT48 project "Humane AI Net" under contract # 952026.
Authors' addresses: S. Mehrotra (corresponding author), C. C. Jorge, and M. L. Tielman, Delft University of Technology, Van Mourik Broekmanweg 6, Delft, The Netherlands, 2628 XE; emails: {s.mehrotra, c.jorge, m.l.tielman}@tudelft.nl; C. M. Jonker, Delft University of Technology & LIACS, Leiden University, Van Mourik Broekmanweg 6, Delft, The Netherlands, 2628 XE; email: c.m.jonker@tudelft.nl.


2160-6455/2023/01-ART4
https://doi.org/10.1145/3610578

## 1 INTRODUCTION

AI technologies are creating new opportunities to improve people's lives worldwide, from healthcare to education to business. However, people do over-trust or under-trust these technologies occasionally [87, 92]. Under-trust can lead to under-reliance, and over-trust can lead to over-compliance, which can negatively impact the task. Hence, for AI systems to reach their potential, people need to have *appropriate* levels of trust in these systems, not just trust. Although there are many ways to define appropriate trust [118], in this article, we take this to mean that the trust a human has in a system needs to align with the actual trustworthiness of the system [32].

It has only been in recent years that we have found research on appropriate trust in AI systems [7, 99, 100, 118]. Appropriate trust is a complex topic, as it requires consideration of the influence of context, the goal-related characteristics of the agent, and the cognitive processes that govern the development and erosion of trust [18]. In this work, we aim to contribute by studying how explanations given by the AI, which highlight different integrity-based principles (e.g., honesty, transparency, fairness), can influence trust and the appropriateness thereof.

**Explainable AI (XAI)** is meant to give insight into the AI's internal model and decision-making [112] and has been shown to help users understand how the system works [16, 85]. Efforts to ensure that AI is trusted appropriately are often in the form of explanations [7, 69, 120]. Intuitively, this makes sense, as understanding an AI system's inner workings and decision-making should, in theory, also allow a user to understand better when to trust or not trust a system to perform a task. Many are focused on how the system works: what it can do and can not [69, 110]. This is done in many different ways, from highlighting essential features of a decision [111], contrasting what would have happened if something was different [91], or how confident the system is about its answer [121].

Typically, explanations are focused on giving information about a system's *ability* to improve appropriate trust. However, literature on how humans trust typically sees trust as more than a belief about ability. Therefore, it is helpful to expand our perspective on explanations as well. A useful starting point for understanding human trust is the **ABI (***Ability, Benevolence, and Integrity***)** model from the organizational context by Mayer et al. [73]. This model has been used extensively in modeling trust, such as by Lee and See [64], Hoffman et al. [39], and Wagner et al. [108]. It defines human trust as "A trusts B if A believes B will act in A's best interest and accept vulnerability to B's action" [73]. Moreover, it distinguishes three trustee characteristics that influence a trustor's trust: belief in ability, benevolence, and integrity.

Ability indicates the skills and competencies to do something. Benevolence is about a willingness to do good to a specific trustor. Integrity is defined as the trustor's perception that the trustee adheres to acceptable principles [73]. One of the extensively studied factors in trust research is the ability of the system [12, 17, 29, 45, 75, 104]. However, fewer studies have investigated the integrity and benevolence dimensions of trust [123]. Benevolence is a specific attachment and emotional connection between the trustor and trustee, which builds over time [73]. Human-agent interactions are often short-term, and the extent to which we form emotional connections needs to be clarified. Therefore, more work on long-term social connections between humans and AI might be necessary before fully understanding the role of benevolence in XAI and human-AI trust relationships.

Prior studies on integrity have linked it to conventional standards of morality—especially those of honesty and fairness [46, 74]. XAI can be regarded as a way to enhance system integrity, i.e., the system being honest about making decisions is a form of integrity. No matter the exact definition, it is clear that integrity is a concept that can play a role even in short-term interactions. Moreover, we follow Huberts in claiming that integrity is an essential concept for human-AI interaction [46].

By applying Olaf's principle,[1] integrity is a necessity and a mandatory requirement of being true to oneself and others [74]. This aligns with the notion that, as AI is increasingly used to make autonomous decisions over time, the principles that underlie these decisions are highly relevant [1]. Furthermore, lack of integrity could cause issues of bias and deception that have already started to impact humankind [62].

Therefore, the question arises what the effect would be of explicitly mentioning principles related to integrity into XAI on appropriate trust of a user in the system. In human-human interactions, principles associated with integrity such as accountability, transparency, and honesty have been suggested as important for appropriate trust [61]. The question arises whether XAI could explicitly use references to these principles in explanations, and how would this affect (the appropriateness of) trust in the system? More specifically, we consider three principles related to integrity to express through explanations:

(1) Honesty about the system's capabilities and confidence.
(2) Transparency about the process of decision-making.
(3) Fairness in terms of sharing what risks such as biases exist.

Honesty, transparency, and fairness appear in various studies as common elements of integrity in HCI, HRI, or human-AI interaction literature [9, 25, 50, 57, 58] (see Section 3). Therefore, in this study, we propose to incorporate references to these principles of integrity in explanations and posit the following research questions:

**RQ1:** How does the expression of different principles of integrity through explanation affect the appropriateness of human's trust in the AI agent?
**RQ2:** How does human trust in the AI agent change given these different expressions of integrity principles?
**RQ3:** How do these different expressions of integrity principles influence the human's decision-making, and do people feel these explanations are useful in making a decision?

We conducted a user study with 160 participants, where they were asked to estimate the calories of different food dishes based on an image of the food with the help of an AI agent. In our user study, the first research question focuses on how different expressions of principles related to integrity (hereafter referred to as "conditions") in explanations can affect appropriate trust in human-AI interaction.

In this article, we study **RQ1** in the context of making an exclusive choice in the form of a decision to choose oneself or an agent to complete the calories estimation task. Moreover, to allow us to study this question, we formally define what it means for trust to be appropriate in this context. **RQ2** aims at understanding change in human trust in the AI agent over time under different expressions of integrity. Finally, **RQ3** helps in understanding the effect of expressions of integrity on human decision-making and the effectiveness of explanations. Additionally, we were interested in exploring possible effects of covariates such as propensity to trust.

***Contributions*** Specially, our research contributes the following:
**1:** We present a measurable construct for appropriate trust in the context of a specific task by providing a formal definition.

[1]McFall [74] describes Olaf's principle as *"An attitude essential to the notion of integrity is that there are some things that one is not prepared to do or some things one must do".*

**2:** We illustrate an approach for expressing integrity of the AI systems with explanations focusing on honesty, transparency, and fairness.

**3:** By conducting a user-study with 160 participants aligned with our research questions, we show how explanations can help in building appropriate human trust in the AI system.

We believe our research holds significance for two main reasons. First, before we can investigate methods to establish suitable trust, it is crucial to have a clear understanding of its meaning. Second, the potential for conveying integrity-related principles through explanations remains largely unexplored. Through our contributions, we aim to broaden our comprehension of fostering appropriate trust between humans and AI, which is vital for effective human-AI interaction [79].

## 2 APPROPRIATE TRUST

### 2.1 Prior Work on Appropriate Trust

To understand what exactly constitutes appropriate trust in Human-AI interaction, we need to understand how people trust each other, i.e., interpersonal trust. Mayer et al. define trust as follows: A trusts B if A believes that B will act in A's best interest and accept vulnerability to B's action [73]. Noteworthy in this definition, and what we believe is a key to defining Human-AI trust, are notions of belief and risk. The interpersonal trust reduces this risk by enabling A's ability to anticipate B, where anticipation is A's belief that B will act in A's best interest. Following Hoffman and Lee and See [39, 64], we carry forward this definition of trust in human-AI interaction.

Recently, there has been rapid progress in studies focusing on building appropriate trust in AI [7, 28, 68, 103, 111, 121]. In a recent work by Yang et al. [118], appropriate trust is defined as the alignment between the perceived and actual performance of the system. This definition talks about the user's ability to rely on the system when correct and recognize when it is incorrect. Similarly, Jorritsma et al. relate appropriate trust to appropriate reliance on the system [53]. On the contrary, Tolmeijer et al. inform us that although both trust and reliance are related, they should be treated and measured as independent concepts [98]. The authors define trust as the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability," while reliance is defined as "a discrete process of engaging or disengaging" based on Lee and See's work [64]. We follow the similar distinction as proposed by Tolmeijer et al. [98] in our work, and see trust as a (subjective) belief, while we see reliance as an (objectively observable) behavior.

Recent works in exploring appropriate trust in human-AI interaction have looked at the role of system trustworthiness and social transparency. For example, Liao and Sundar emphasize the mediating role of information display on trust judgments, and that appropriate trust relies on effective communication of system trustworthiness [68]. However, Ehsan et al. show that social transparency could support forming appropriate trust in human-AI interactions by embedding socio-organizational context into explaining AI-mediated decision-making [28]. Additionally, various works in human-robot interaction focus on providing end-users with an accurate mental model of a robot's capabilities for establishing an appropriate level of trust [23, 56, 84, 107]. We believe that in the above-mentioned prior works, the provided constructs of appropriate trust are limited. The majority of these works consider the system's ability or performance for defining appropriate trust. We would argue that there is more to appropriate trust than a correct belief in the ability of the system; such as the psychology of trust focusing on beliefs [67], mutualistic benevolence impacting trust [63], personal integrity requiring truth telling [74], and ethics of trust focusing on fairness [71], or even environment-based factors including task and culture [94].

A substantial amount of literature in human-AI interaction focuses on calibrating human trust, which is the process of making trust more appropriate over time. For example, De Visser et al. defined trust calibration based on prior works by Cohen et al. and Lee and See [20, 64] as a process

of updating the trust stance by aligning the perception of an actor's trustworthiness with its actual trustworthiness [23]. According to them, calibrated trust is a function of perceived trustworthiness that helps in eliciting appropriateness of trust. Okamura and Yamada proposed a framework for detecting inappropriate trust in a system with a behavior-based approach [83]. Their framework detects over- and under-trust in the system by monitoring the user's reliance behavior. In a similar work by McGuirl and Sarter, the AI system provided system confidence information to improve trust calibration [75]. In the above-mentioned studies, the focus of the task was to calibrate human trust. These related works can be helpful to understand the appropriateness of trust, as calibration is about the process that incorporates updating trust levels, and appropriate trust can be Boolean per situation resulting from that update.

In other works by Mehrotra et al. and Winikoff [80, 116], it has been argued that AI systems' value-based reasoning can help achieve appropriate trust. Mehrotra et al. showed the effect of (dis)-similarity of human and agent's values on a human's trust, which forms a part of appropriate trust [80]. Similarly, in work by Winikoff [116], we can find theoretical foundations for achieving appropriate trust based on value-based reasoning. According to Winikoff [116], value-based reasoning is an essential prerequisite for human-AI interaction, because (a) an AI system that is able to conduct reasoning using human values to make decisions could be used as a basis for providing higher-level and more human-oriented support (b) having an explicit model of values can help in verifying AI system's behavior, for example, in system's reasoning and decision-making by taking ethical considerations into account. Building on these works, our research looks for a deeper understanding in evaluating appropriate trust in human-AI interaction by incorporating integrity-based explanations where integrity in itself is a part of basic inherent human values.

### 2.2 Our Approach on Appropriate Trust—a Formal Perspective

We are interested in the effect of integrity-based explanations on appropriate trust, so we need to first understand what exactly appropriate trust is and what counts as over- or under-trust. Over-trust is often related to over-reliance on the system leading to misuse, and under-trust is related to under-reliance on the system leading to disuse. Also, we define another trust category—inconsistency—following Sadiku et al. [93], who quotes famous anthropologist Margaret Mead on understanding psychological notions of human behavior: "*What people say, what people do, and what they say they do are entirely different things.*" Intuitively, inconsistency happens when people choose to rely on those they trust less, or vice versa.

The work described in the previous paragraph provides a conceptual understanding of appropriate trust, which we build on. Most notably, we say appropriate trust occurs when a belief about trustworthiness matches with actual trustworthiness. We consider appropriate trust as a state that is either true or false, rather than looking at the whole calibration process. However, for our purposes, we also require a practically measurable definition of trust on top of this conceptual understanding. Therefore, we propose a formal definition that tells us exactly in which situations trust is appropriate or not. Specifically, we consider appropriate trust from a specific angle in this article. Our definition does not try to give an all-encompassing definition of appropriate trust, but rather does so in the context of a specific type of task. Namely: Our task involves an exclusive choice of who will perform the task, the agent or the human. This selection is motivated by prior works on choice behavior by Israelsen and Ahmed and Okumara et al. [48, 83] and recent work by Miller [82]. During our task, a user and an AI agent are working jointly. **The user should select whether for a particular task they want to rely on the AI agent or do it themselves**. In this situation, we define trustworthiness of the agent as how well they perform this task.

In our definitions, we use $TW$ for describing trustworthiness. When discussing the trust of a human $h$ in an AI agent $a$ for a task $t$, we do not write $T_h(a, t)$ but $T_{(human \rightarrow agent)}$, dropping the $t$

for the ease of reading. We then define:

$$T_{(human \rightarrow agent)} = \text{Trust of the human in the agent for accomplishing a task}$$

$$TW_{human} = \text{(actual) Trustworthiness of the human for a task}$$

$$B_{human}(TW_{human}) = \text{Belief of the human regarding its own Trustworthiness for a task}$$

$$TW_{agent} = \text{(actual) Trustworthiness of the agent for a task}$$

$$Selection_{human} = \text{Selection by the human for the task, i.e., themselves or the agent}$$

We define appropriate trust based on the action and the subjective opinion of the human, as well as the trustworthiness of both human and agent. Now, we will describe our concepts with the help of above-mentioned parameters:

- **Appropriate Trust:** *(a)* the human estimates that the AI agent is better at the task than the human, *(b)* also the actual TW of the AI agent is equal to or higher than the human's TW, and *(c)* the human selects the AI agent for the task and *vice versa*—Equations (1) and (2). Here, *(a)* is cognitive trust from the human, *(b)* is god's eye view of the *TW* (described in the next section) and *(c)* is human selection that could be based on observable behavior, rationality or simply delegation of the responsibility.

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} \leq TW_{agent}] \wedge Selection_{human} = agent \quad (1)$$

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} \geq TW_{agent}] \wedge Selection_{human} = human \quad (2)$$

- **Over-trust in the agent:** the human estimates that the AI agent is better at the task than the human and selects the AI agent even though the actual TW of the AI agent is lower than the human's TW.

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} > TW_{agent}] \wedge Selection_{human} = agent \quad (3)$$

- **Under-trust in the agent:** the human estimates that they are better at the task than the AI agent and select themselves even though the actual TW of the AI agent is higher than the human's TW.

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} < TW_{agent}] \wedge Selection_{human} = human \quad (4)$$

There could be instances where one can trust someone more than themselves and still choose not to rely on them and vice versa. For example, we rarely doubt the efficacy of automatic shifting mechanisms of today's cars, yet some people still choose to manually shift for the pleasure of it. However, people might want to avoid responsibility by delegating to the other person even if they have higher trust in themselves. Therefore, we formulate two additional cases based as:

- **Inconsistency with a good outcome:** the human estimates that the they are better at the task than the AI agent however, they select the agent for the task, and the actual TW of the AI agent is higher (or equal) than the human's TW and vice versa.

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} \leq TW_{agent}] \wedge Selection_{human} = agent \quad (5)$$

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} \geq TW_{agent}] \wedge Selection_{human} = human \quad (6)$$

Table 1. Categorization of the Trust Categories Based on Equations (1) to (8)

| Equation | Higher TW | Human trusts who? | Human selects | Trust Category |
|---|---|---|---|---|
| 1 | AI Agent | AI Agent | AI Agent | Appropriate |
| 1 | Equal | AI Agent | AI Agent | Appropriate |
| 2 | Human | Human | Human | Appropriate |
| 2 | Equal | Human | Human | Appropriate |
| 3 | Human | AI Agent | AI Agent | Over-trust |
| 4 | AI Agent | Human | Human | Under-trust |
| 5 | Human | AI Agent | Human | Inconsistency with a good outcome |
| 5 | Equal | Human | AI Agent | Inconsistency with a good outcome |
| 6 | AI Agent | Human | AI Agent | Inconsistency with a good outcome |
| 6 | Equal | AI Agent | Human | Inconsistency with a good outcome |
| 7 | AI Agent | AI Agent | Human | Inconsistency with a bad outcome |
| 8 | Human | Human | AI Agent | Inconsistency with a bad outcome |

- **Inconsistency with a bad outcome:** the human estimates that the AI agent is better at the task than the human, however, they select themselves, but the actual TW of the AI agent is higher than the human's TW and vice versa.

$$[T_{(human \rightarrow agent)} > B_{human}(TW_{human})] \wedge [TW_{human} < TW_{agent}] \wedge Selection_{human} = human \quad (7)$$

$$[T_{(human \rightarrow agent)} < B_{human}(TW_{human})] \wedge [TW_{human} > TW_{agent}] \wedge Selection_{human} = agent \quad (8)$$

Our definitions are suited for our task requiring exclusive decision-making, i.e., tasks where one has to make a decision by either relying on oneself or the other party. We now summarize the cases mentioned above in the following table. In Table 1, Equation (1) represents two conditions where $\mathrm{TW}_{human} = \mathrm{TW}_{agent}$ and $\mathrm{TW}_{human} < \mathrm{TW}_{agent}$ keeping other comparisons same. A similar pattern follows for Equations (2), (5), and (6).

## 3 INTEGRITY

### 3.1 Prior Work on Integrity

Mayer et al. state that "the relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable" [73]. This definition of integrity is rooted in the studies on organizational management. However, definitions of integrity vary across disciplines, but even within disciplines. For example, again in management science, according to Jeavons [49], integrity has to do with continuity between appearance and reality, intention and action, promise and performance, in every aspect of a person's or an organization's existence; whereas Hon and J. E. Grunig [42] described integrity in public relations as ''the belief that the other party is fair and just.''

A literature review by Palanski [86] provides an overview of relevant integrity definitions in philosophy. The review outlines five general categories of integrity: wholeness, consistency of words and actions, consistency in adversity, being true to oneself, and moral/ethical behavior. Other research in human communication research measures integrity by simply "being honest" or "having integrity" [117]. Turning to integrity in human-computer interaction, we see similar concepts taking the form of integrity definitions. For example, McKnight defined integrity as beliefs of honesty and promise-keeping for building trust in e-commerce systems [76]. Jensen et al. measured the integrity of a drone system as being truthful in communication, honest, keeping commitments, being sincere and genuine, and performing as expected [50]. In both the studies mentioned above

and in References [90, 102], integrity in human-AI interaction is strongly related to honesty and being transparent about the process of decision-making.

Kim et al. [57] and Wang and Benbasat [9] explored integrity in terms of fair dealings and unbiased decision-making approaches. Kim et al. found that a robot's integrity is responsible for mediating the relationships between a robot and human trustworthiness. In a recent work by Knowles and Richards, integrity is highlighted in promoting public trust in AI [58]. According to the authors, trust in AI arises in part from a perception of coherence between the human norms as highlighted by Giddens [37]. Giddens talks about human norms in two dimensions—the degree to which agents within the institution are empowered and the use of language by the AI agents. These dimensions resonate with scholarship on trust that emphasizes the importance of integrity. In summary, we can understand that integrity of an AI agent plays an essential role in building trust. Some approaches link integrity to the sharing of (moral) principles or keeping to human norms. In other approaches, specific principles are mentioned to constitute integrity. Differences exist, but some common principles related to integrity are honesty, keeping promises/commitments, consistency, and fairness. For AI in particular, transparency in decision-making is often mentioned as key to integrity as well.

### 3.2 Our Approach on Integrity: Expressions of Integrity through Explanations

As discussed in the previous section, many definitions of integrity in the AI literature focus on specific principles. In this article, we specifically focus on three of them: honesty, transparency, and fairness. These are all often used as honesty [50, 66, 117], transparency [5, 25], fairness [15, 58]. Although keeping commitments [97] and consistency [88] are also often used, we choose not to use them in our setting. Keeping commitments and being consistent both imply longer-term interaction and would be most logically related to behaviors more than explanations. Moreover, we could imagine more principles of integrity are used in different settings. We do not argue our list is complete, but rather make a starting point with three important principles to potentially incorporate in XAI.

Honesty, transparency, and fairness are all complex concepts that should be employed in decision-making of AI [3]. In this article, we choose to express elements of honesty, transparency, and fairness in a way that suits XAI. This means we do not claim that our explanations fit the full picture of what it means for an AI system to be, e.g., "honest." Rather, we designed a specific set of explanations aimed at highlighting: honesty in terms of highlighting uncertainty and confidence; transparency in terms of explaining the process of decision-making; and fairness in terms of sharing with users the possible risks and biases that may exist in the advice.

We picked these specifications, as they make sense for AI to use in explanations. Uncertainty is often highlighted in confidence explanations [101], transparency is often mentioned as a keystone of AI, and the decision-making process is something that should be particularly transparent [31], and giving fair advice means not only trying to exclude biases and risks as much as possible, but also being open about this [78]. These specifications also align with the work of Wang and Yin, who provided three desiderata of designing effective AI explanations [112]. These desiderata include (a) designing explanations improve people's understanding of the AI model, (b) helping people recognize the uncertainty underlying an AI prediction, and (c) empowering people to trust the AI appropriately. For brevity's sake, we will use the broader terms "honesty," "transparency," and "fairness about risks" to refer to our specific expressions in the remainder of this article.

*3.2.1 Design of Explanations.* Based on these specifications, two researchers with a Computer Science background and one with a Cognitive Science background brainstormed together and generated sentences that formed explanations expressing the principles of integrity in three different

ways. We followed the notion of situation vignettes following the work by Strackand and Gennerich [96] to create text-based explanations.

Each explanation creator was provided with a stack of different expressions of the principles of integrity as identified above. Each note card had one expression printed on it. Each creator read through each other's explanations and decided if they felt it fell within scope or out of scope of the principle to be expressed. For each explanation, creators then described their reasoning for classifying the expression of integrity as within or out of scope. In the end, all creators engaged in similar reconstructive processes to finalize the explanations by controlling the length (word limit [6]) of explanations for the three integrity aspects (honesty, transparency, and fairness). Overall, three iterations were performed for each explanation. The main author followed up with any necessary questions to determine the researcher's interpretation of each hypothetical situation.

Once the explanations were completed, we divided them in a four-part schema. We chose to follow a schema to keep consistency and uniformity in the integrity-based explanations throughout different conditions. Also, keeping a schema supports designing AI agents who can provide forward-reasoning decision support [122], i.e., helping people understand the information in phases and make an informed decision.

The first part of the schema shows an explicit reference to the integrity principle, for example, *I think it is important to be transparent, so I'll tell you how I came to this decision.* This means that the agent explicitly acknowledges that they value a certain principle. Further, all explanations contain a reference to the source of the data on which the suggestion is based; an estimation of the total calorie count based on the identified ingredients; and the answer the agent picks.

To compare the different expressions of integrity, a baseline explanation was also designed. This type of explanation did not include a specific reference to an integrity principle, but always expressed the source of the data, an estimation of calories without referencing the ingredients, and the final answer chosen by the AI agent.

*3.2.2 Expressions of Integrity in Explanations.* Our expressions of integrity are portrayed based on the following schemas. In addition to elements of the baseline explanation (the source of the ingredients and the final answer), all the integrity-based explanations included a list of ingredients identified by the AI system of a food plate. Variation was added to avoid mechanical and "fake"-looking explanations. Specific examples of the different ways of expressions of integrity through explanations can be found in Table 2 and in the supplementary folder.

(1) **Honesty** explanations always start with a reference to honesty, followed by an estimation of how sure about the total calories on the plate (e.g., "so I'll tell you that I'm not entirely sure about identifying the total calories on this plate"). Often a confidence % is already added, and usually there are at least two statements (e.g., one explaining why this confidence level, one giving options on what the dish could be, or what it could contain, e.g., "It could be Caprese salad with 88% of confidence or beet salad with 85% of confidence."
(2) **Transparency** explanations always start with a reference to transparency, followed directly by the selected answer and a ''I'll tell you how I came to this decision." Following (usually directly) is an indication of how sure the system is of what it could possibly be, sometimes in combination (e.g., "I'm almost sure this is x, however, I'm not sure about item x"). Sometimes there is a further explanation of why the system is this sure (e.g., "because of the low image resolution," "My algorithm has failed to recognize the identified portion"), or some more information about the dish (e.g., "Salsa is usually spooned over nachos and are sprinkled with grated mozzarella").
(3) **Fairness about risks** explanations always start with a reference to bias, followed by an indication of how sure the system is of what it could possibly be, sometimes in

Table 2. Different Ways of Expressing Integrity through Explanation by an AI Agent

| Expression of integrity | Explanation |
|---|---|
| `Baseline` (Average length = 55 words, SD = 6 words) | *The ingredients that I can correctly identify are displayed in the list and their confidence scores. The information I have is based on the UNESCO food nutrition website data. On adding, the total calorie count is 738 calories. Therefore, I would tick option 750 based on the identified ingredients.* |
| `Honesty` (Average length = 125 words, SD = 23 words) | *I think it is important to be **honest**, so I'll tell you that I'm not entirely sure about identifying the total calories on this plate. I am not confident about the food item encircled in dark white circle. This is because I have limited training data matching with this encircled food item. The items that I correctly identified are in the table. The information I have is based on the data taken from UNESCO food nutrition website. On adding, the total count is 750 calories which is closer to 738. Therefore, I would tick the option 750 with my overall confidence level as 62.5%. This confidence level means I am moderately sure about my answer.* |
| `Transparency` (Average length = 128 words, SD = 19 words) | *I have selected 750 calories as the answer to this question. I think it is important to be **transparent**, so I'll tell you how I came to this decision. I found a similar dish based on my training data from the UNESCO Food & Nutrition website that closely matches the given plate for the calories count. The dish I found is a curry; however, I am not sure about which curry it is. The matching visualization is shown next to the identified ingredients. Based on my training data and similar dish search, the total amount of calories should be 738 calories with 62.5% confidence, similar to the best match example.* |
| `Fairness about risk` (Average length = 130 words, SD = 25 words) | *I think it is important to be **fair and unbiased**, so I will explain how I combat bias in my answer. I'm not entirely sure about identifying the total calories on this plate. I am not confident about the food item encircled in dark white circle. This is because there is no clear pattern among human annotators of this image. This image is labelled as an Indian Madras curry from UNESCO food nutrition website but I can find annotators for its ingredients only from the western population out of which no one has a profession tag of chef. They have classified the encircled item as bay leaf, fish, meat, chicken or beef. The items that I correctly identified are in the table which gives an estimate of 738 calories. Combining all the existing knowledge with uncertainty regarding the encircled item I will select the option 750 with my overall confidence level as 62.5%.* |

combination (e.g., "I'm almost sure this is x, however, I'm not sure about item x"). The reasoning explanation can be an explanation of the (lack of) confidence for a choice or of the choice itself. There is always either an explanation of the confidence or an explicit reference to how large the chance of bias in the process or data would be or even both. In some cases, there is a warning with the final answer that bias might be present. The specific explanation is unique for every dish, so no explanation is repeated exactly.

#### 3.2.3 *Differences between Integrity-based Explanations.*

(1) **Baseline vs. Integrity Conditions:** The baseline lacks a reference to any specific principle and only refers to the source of the data used, an estimation of the total calorie count, and the final answer. The three integrity conditions all include this data source, estimation, and answer as well. In addition, they each explicitly refer to their own principle to start.
(2) **Honesty vs. Transparency and Fairness:** Honesty explanations prioritize providing accurate and truthful information about the AI agent's decision-making process and highlighting uncertainty. Also, it is the only one that explains what the confidence intervals mean.

(3) **Transparency vs. Honesty and Fairness:** Transparency explanations aim to provide a comprehensive and understandable view of the AI agent's inner workings, without necessarily prioritizing the accuracy or truthfulness of the information provided. Also, it is the only one with a visual representation of ingredients identified, includes more references to what the decision is based on, and mentions the final decision both at the start and end, rather than just the end.
(4) **Fairness vs. Honesty and Transparency:** Fairness explanations focus on ensuring that the AI agent's decision-making process does not unfairly discriminate against certain individuals or groups. It also explains why it is certain and where biases might occur more than the others.

We also designed visual explanations exclusively for the transparency condition of the integrity, as this notion deals with the process of decision-making. Our classifier provided comparative examples of visual classification. These visualizations categorize confidence values into buckets, such as High/Medium/Low, showing the category rather than the numerical value. The cutoff points for the categories were best match (confidence score $> 0.8$), good match ($0.5 <$ confidence score $< 0.79$), and unsure match (confidence score $< 0.49$); refer to Figure 9. These cutoff points were set in accordance with a prior study by Kocielnik et al. [59] and Google's PAIR guidebook [2].

## 4 METHOD

### 4.1 Participants

One-hundred-eighty-two participants (89 female, 93 male) were recruited to participate in the study, via the online crowdsourcing platform Prolific (mean age = 24.8 years, SD = 4.4 years) and the student university mailing list (mean age = 22.1 years, SD = 2.3 years). We recruited through two different methods, because we had less turnout of students from the mailing list due to long study completion time. There were no differences among the two samples of participants for the responses we received.

A total of 121 participants participated through the crowdsourcing platform and 61 through the university mailing list. We chose Prolific platform because it is an effective and reliable choice for running relatively complex and time-consuming interactive information retrieval studies [99]. Participants were selected based on the following criteria: age range (18+ years old); fluent level of English—to ensure that participants could understand the instructions; and had no eating disorder—to ensure minimal risk to participants for viewing different food items.

Thirty-five percent of the participants reported having studied computer science or some related field. Our participants were from 30 different countries, with most participants reportedly born in the United Kingdom (35), Germany (26), the USA (20), and India (20). Participants were informed about the nature of the task and the total completion of around 35 minutes. Those who accepted our task received brief instructions about the task and were asked to sign an informed consent before beginning their task session.

The study was approved by the Human Research Ethics Review Board of Delft University of Technology (IRB #2021-1779). Prolific participants received an honorarium of £ 5.43/hr for their participation. All participants were provided an option to participate in 5x 15 Euro Amazon gift voucher raffle prize.

### 4.2 Task Design

We aimed to establish *human-in-the-loop* collaboration in our experiment; i.e., a human making a decision with the assistance of an AI assistant. In our experiment, participants were asked to estimate the calories of different food dishes based on an image of the food. We designed this task around calories as an approachable domain for our participants. The food dishes in our experiment

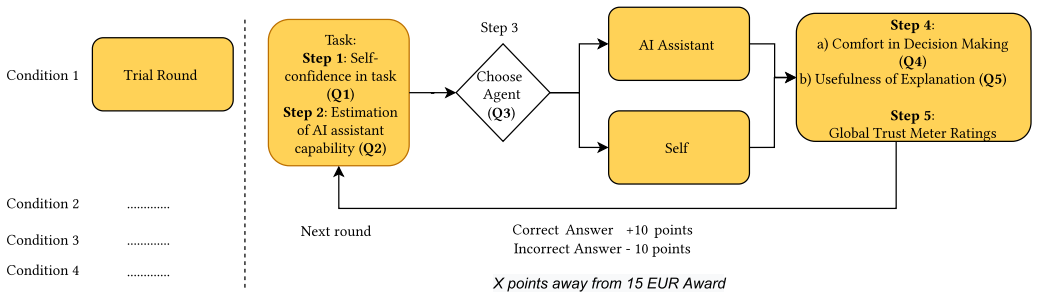


Fig. 1. A between-subject measure of demonstrated trust. Participants interact with an agent and then must choose whether to rely on themselves or the AI assistant to complete each task in a sequence of tasks. An incorrect answer is a risk to the trustor causing reduction of 10 points and further away from the required points to receive an award.

were specialized dishes from different countries around the globe. It is rare that participants can judge all the food dishes well but are often good at judging their own cuisine. Therefore, we told participants that there is an AI assistant to help them in identifying the correct amount of calories.

During the brainstorming session of the authors, we decided to use the Food-pics database [11] for selecting our dishes. We selected this database because it contains most popular dishes for European and North-American populations from across the globe along with detailed meta-data of the dishes. Fifteen randomly selected food dishes (referred to as ''rounds'' hereafter) were taken from this database in the main experiment. Each round consisted of five steps.

**Steps of the task:** At the *first* step, participants were shown an image of a food dish. They were asked to select their confidence in correctly estimating the calories of the food dish. Specifically, we asked our participants, on a scale of 1–10, with 1 being ''Not at all confident'' and 10 being ''Fully confident,'' How accurately can you estimate the calories of this food image (Q1)? A zoom-in option was also provided to participants to have a closer look at different ingredients of the food image. Subsequently, they were asked to guess one of the four options they believed to be closest to the correct amount of calories in the dish. One option out of four was always the correct answer, and the first step only involved guessing the correct answer.

At the *second* step, an AI assistant guessed the correct answer from the same options as step one. The AI assistant provided a list of ingredients that it believed to be a part of the dish and the dish name with confidence scores (*for details, refer to Figure 3*) in real time. The AI assistant also explained the reasoning for an answer by providing explanations. Additionally, at this step, participants were asked (Q2) to tick a checkbox if they believed that the AI assistant could better estimate the calories than themselves. At the *third* step, participants selected their final decision by choosing between themselves or the AI assistant (Q3). At the *fourth* step, participants rated their comfort level in making the decision (Q4) and usefulness of explanations (Q5). Finally, at the *fifth* step, the correct answer was shown to the participants and participants were asked to adjust their trust level in the AI assistant. An overview of the above steps is visualized in Figure 1.

**Scoring method:** Each correct answer yielded +10 points, and an incorrect answer cost -10 points. We specifically applied -10 points for a wrong answer to involve the risk factor associated with trust. Additionally, participants were informed that if they end up in the top three scorers on the leaderboard, they will qualify to receive a 15 Euro gift voucher. The idea to include the leaderboard was to turn a single-player experience into a social competition and provide participants with a clear goal. Participants were only informed about the top scores of the leaderboard and

their rank once they finished the task. We did this to ensure that participants make an informed selection till the end of the task to qualify for the prize. Based on our exit interviews, participants were careful with their selection, as they wanted to maximize their chance of winning the award.

### 4.3 Measures

We used two types of measures. First, subjective measures where users directly report their opinion (referred to as "subjective measurement" hereafter) (e.g., References [21, 36, 119]). Second, behavioral measures (e.g., reliance [14, 27] and trustworthiness, e.g., References [32, 34, 50]). We used the wording AI assistant instead of AI agent for the ease of participants.

**Subjective measures:** Guided by the trust definition in the human communication research domain [114], we measured participant's trust inspired by Yang et al. [118] as four different measures: (1) cognitive trust to understand human estimation of AI agent capabilities [52], (2) participant's comfort level in making a decision [118], (3) usefulness of the AI assistant explanation [118], and (4) a global trust meter that captures changes in trust [55].

*First*, human cognitive trust to follow the AI assistant recommendation was measured via Q2: "Select this [check] box if you think that the AI agent can better estimate the calories than yourself." We informed our participants that by selecting the check-box they believe that the AI agent is better at the task than themselves.

*Second*, human comfort was measured by the question: Q4—"How do you feel about your decision?" this question measured participants' comfort in taking a decision and was rated on a 10-point Likert scale from *Not at all comfortable (1) to Very comfortable (10)* with a step size of 0.2, i.e., step sizes were 1.0, 1.2, 1.4...9.8, 10.0. We included this question in our user study for two reasons: (1) based on recent work by Yang et al. [118] indicating the importance of human comfort in decision-making and (2) based on our pilot study where participants often used the word "comfortable" to describe their decision, which also matches with prior work by Wangberg and Muchinsky [109].

*Third*, AI assistant explanation was measured by the question Q5: "Was the explanation by the AI assistant helpful in making the decision?" This item was rated on a 10-point Likert scale from *Not at all helpful (1) to Very helpful (10)* with a step size of 0.2.

*Finally*, a linear "Trust Meter" ranged from complete distrust (0) to complete trust (+100), inspired by Khasawneh et al. [55]. Participants were asked to adjust the trust meter after every round if their trust in the AI assistant changes. The trust meter was always available to participants and took the previous round's trust meter value in every new round. For the first round, the default value of the trust meter was set at 50.

**Behavioral measures:** For trustworthiness and reliance on the system, we looked at what the participant and AI agent did. *First*, our **trustworthiness (TW)** measurement was about who was better at the task, so could be either the participant, the AI agent, or both. It was measured by considering how far the selected option was from the correct answer. No two options among the four options were equal distance from each other. For example, if available options are 25, 66, 97, and 143, of which the correct answer is 97, and human selection is 66 and AI agent selection is 143, then human TW is higher than the AI.

*Second*, participants were asked to "*Select your final decision by selecting among the two options—yourself or the AI assistant's guess*" (Q3). With Q3, we measured reliance (distinct from trust, as we discussed in the introduction) by analyzing the behavior of the participants. If they followed the AI assistant's advice or decision and selected it, then they were considered to rely on it. If they switched their answer to another answer than the advised answer, then they did not. In case the two options were same, participants were asked to still decide based on the reasoning for calories of

the dish, classification of ingredients, and confidence levels. Their choice determined their reliance behavior on the AI agent.

It is important to note that although trust and reliance are related concepts, they should be measured as independent concepts. In this work, we follow this distinction as pronounced by Tolmeijer et al. [98], where trust is the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [64, p. 51], while reliance is "a discrete process of engaging or disengaging" [64, p. 50] with the AI agent.

### 4.4 Experimental Setup

The study was a mixed between- and within-subject design. The within-subject factor was subjective ratings and between-subject factor was the integrity condition. This design choice was inspired by Hussein et al. [47]. Participants were randomly assigned to one of four different experimental conditions ("Baseline," "Honesty," "Transparency," and "Fairness"). Each condition had an equal number of participants. We did not manipulate other factors such as time [81] and workload [22], but we controlled reliability [65] and risk factors [77]. The advantage of this experimental setup, as stated by Miller [82], is that we can perform detailed analysis on the relationship *Trustworthiness* → *Perceived Trust*, which in turn helps in understanding appropriate trust.

We utilized Clarifai Predict API with the "Food" model to recognize food items in images down to the ingredient level.[2] Our visual classifier returned a list of concepts (such as specific food items and visible ingredients) with corresponding probability scores on the likelihood that these concepts are contained within the image. Our pre-trained classifier accuracy was about 75% (11/15 = 73.33%), roughly matching the average actual classifier's accuracy of 72%. The list of ingredients along with their confidence score was represented in the form of a table, as shown in Figure 3.

**Sequence of trials:** Each participant finished all 15 rounds, including a trial round. The number of rounds was decided to (1) compare with other experiments that studied trust (e.g., References [99, 118]), (2) have enough trials to develop trust but prevent participants from memorizing the order (serial position effects [44]), and (3) have sufficient data for all the integrity conditions.

In each condition, participants finished a sequence of trials. All the sequences had the identical order of correct/incorrect recommendations by the AI assistant. This identical order allowed us to compare different conditions. We also ensured that the AI agent response in the trial round was always correct to protect trust in an early stage and to not skew or strongly bias towards wrong [72]. Food dishes in the sequence were randomized, and the instances used for training and practice were excluded in the main trials. On completion, participants were asked to fill in a post-experiment questionnaire targeted towards (a) their overall experience, (b) possible reasons for their changes in trust meter, and (c) their decision to select themselves or the AI assistant.

**Pilot Study and Pre-test of Explanations:** We used a think-aloud protocol with three participants for a pilot study. The aim of the pilot study was to test the experiment design and check the explanations manipulations. In our experiment, participants were comfortable with estimating calories of the food dishes based on their familiarity with the cuisine and often chose the AI agent when they were not confident. For example, a participant who identified himself as an American often relied on the AI agent to guess a food dish from Myanmar. Similarly, another participant who identified herself as an Asian often relied on the AI agent for a Mexican food dish. Based on these observations and UI layout feedback from the participants, we fine-tuned the questions and instructions. After the experiment was finished, we checked for manipulation of explanations. We asked our participants to describe the principle of integrity they saw in the experiment from the

[2] https://www.clarifai.com/models/ai-food-recognition.

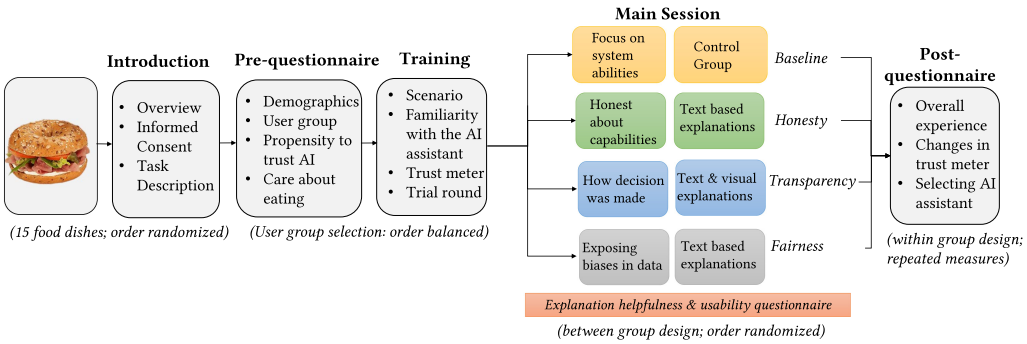


Fig. 2. This figure illustrates the experimental design of the user study. Each participant was assigned to a experimental condition (Baseline, Honesty, Transparency, and Fairness about risk), and they finished 15 rounds in approximately 35 minutes with a 2-minute break after 7 rounds to avoid fatigue effect [19].

note cards that we used earlier with the explanation creators. All participants correctly identified the integrity principles from the note cards. This result helped us in pre-testing our explanation and start with the main experiment. We excluded these three participants from the main experiment.

### 4.5 Procedure

After participants provided informed consent, they saw an overview of the experiment. As shown in Figure 2, participants were first asked to complete a pre-task questionnaire consisting of (i) demographic questions about their age and gender as well as (ii) the propensity to trust scale [35] (Q6) and a balanced diet eating question (Q7) on a 10-point Likert scale from *"I don't care of what I eat" to "I care a lot of what I eat."*

At the beginning of the experiment, we told participants that they would work with an AI assistant and hinted that it could be wrong in its recommendation. They then took part in a trial session, read the instructions, saw an example of a food dish, and practiced using the trust meter. Participants then proceeded to the main session. For each *step,* as explained in Section 4.3, participants first saw an introduction of what they could expect to see. In addition, they were asked to focus on the table generated by the AI assistant for specific food items and visible ingredients with corresponding probability scores. The screenshots of each step are in Figure 3.

## 5 RESULTS

One-hundred-eighty-two participants participated in the user study, of which 19 (18 from Prolific and one from the university mailing list) did not pass our attention checks, leaving us with 163 participants. Furthermore, one participant selected the AI agent, and two always selected themselves, with a total experiment time of only eight minutes, indicating potentially invalid data. Hence, we removed the data of those three participants. Thus, the results and analysis include the remaining (160 participants (female = 85, male = 75; mean age = 23.6 years, SD = 2.8 years). A power analysis of the mixed ANOVA with G*Power tool [30] revealed that with 40 participants per group, we have a power of 0.93 (considering a medium effect size of $f = 0.25$, $\alpha_{new} < .046$).

### 5.1 Effect of Different Principles of Integrity on Appropriate Trust

In this subsection, we analyzed how the expression of different principles of integrity through explanation affects appropriateness of the trust of a human in that agent (**RQ1**). For this analysis, we first conducted a descriptive statistics and then performed inferential statistics on the collected

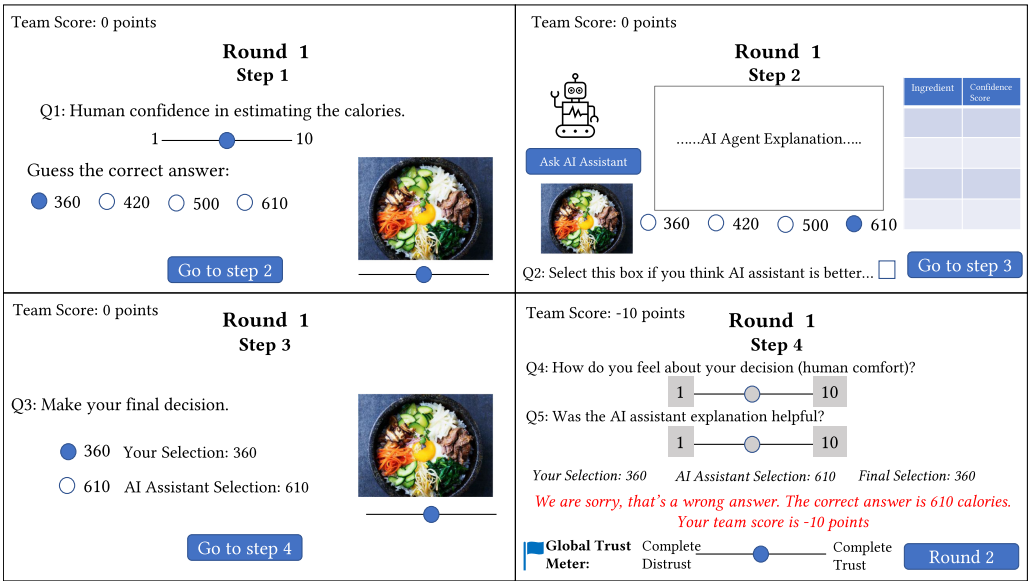


Fig. 3. Illustration (a simplified version) of the four steps performed by a participant of the user study. In step 1, participants rate their confidence in accurately identifying the calories (Q1). In step 2, the AI agent selects its answer with its reasoning in the form of explanations and confidence scores (Q2). In step 3, the participants makes their final decision (Q3). Finally, in step 4, participants rate their comfort in decision-making and usefulness of the explanations (Q4 and Q5).

Table 3. A Contingency Table of Frequency Distribution Illustrating Number of Times Different Trust Categories Were Observed Given Explanations Highlighting Different Principles of Integrity

| Condition | App. Trust | Inconsistency(Bad) | Inconsistency(Good) | Under-trust | Over-trust |
|---|---|---|---|---|---|
| Baseline | 0.418 | 0.078 | 0.327 | 0.123 | 0.050 |
| Honesty | 0.433 | 0.068 | 0.285 | 0.158 | 0.053 |
| Transparency | 0.410 | 0.068 | 0.302 | 0.153 | 0.065 |
| Fairness | 0.552 | 0.060 | 0.218 | 0.088 | 0.088 |

Occurrences are scaled as % distributions between 0 (no occurrence) -1 (always occurred).

data to study the effect of explanations. The post-experiment questionnaire responses were analyzed to support the results and are reported in Section 6.1.

The categorization of trust categories was calculated based on Table 1. Following the equations in the table, Higher TW was derived based on the TW measurement (as described in Section 4.3). The value for Human trusts who? was based on the participant's response for Q2, and for Human selection, it was based on Q3. On entering these values in Table 1, we got our five different trust categories, as described in Section 2.2.

**Frequency Distribution:** Table 3 shows the frequency distribution of different trust categories as observed for the explanations expressing different principles of integrity. For example, consider a participant who viewed explanations expressing honesty about uncertainty and who fell into the appropriate trust category seven times, inconsistency (good and bad outcome) two times each, under-trust three times, and over-trust once. Then, for the honesty condition in Table 3, we report appropriate trust as 0.46, inconsistency (good and bad outcome) as 0.13 each, over-trust as 0.20, and

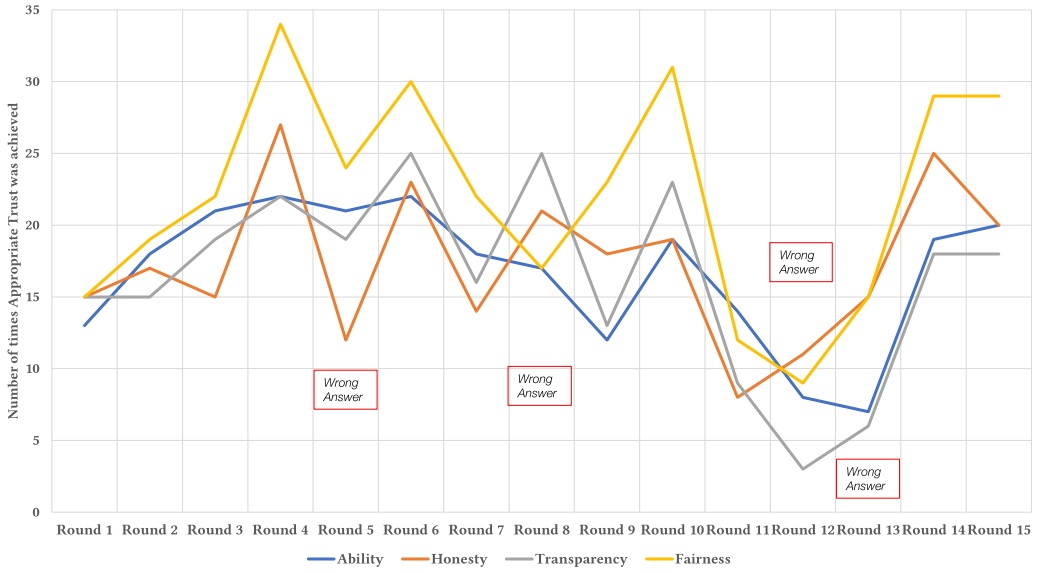


Fig. 4. This figure illustrates the frequency distribution of appropriate trust across 15 rounds and how it is affected by the wrong answers.

under-trust as 0.06 on a scale of 0–1. Each condition consists of data of 40 participants collected over 15 rounds, i.e., 600 data points per condition.

**Effect of Integrity Expressions:** We found a statistically significant effect of the integrity principles expressed through explanation on trust categories. A chi-square test of independence $\chi^2(12, N = 40) = 55.11$, $p < .001$, $\varphi_c = 0.30$ showed that there is a significant relationship between trust categories and experimental conditions. We further analyzed our contingency table (*Table* 3) as a mosaic plot [40] to investigate relationships between different trust categories and conditions. While constructing the mosaic plot, we extracted Pearson residuals from the output of the $\chi^2$ results.

We visualized Pearson residuals contribution to the total chi-square score using the correlation plot (*for details, refer to Appendix A, Figure 6*) as our exploratory analysis. Following the correlation plot, a correlation value of $\rho = 3.45$ between the "Fairness about risk" explanation and appropriate trust category was found. Following Hong and Oh [43], this correlation implies a strong association between the "Fairness about risk" explanation and the appropriate trust category.

We were also interested in understanding how different trust categories build up or are relatively stable over time and how they are affected by the wrong answer. Figure 4 illustrates the frequency distribution of appropriate trust across 15 rounds. The figure shows that appropriate trust drops with the first wrong answer across the four conditions. However, this effect does not perpetuate in later rounds. It is interesting to note that appropriate trust builds up over time (rounds 1 to 4) and recovers slowly after each wrong answer. We also provide a similar graph as Figure 4 in the supplementary for other trust categories.

**Predictors for Trust Categories:** The trust categories were binary variables in our study: Either the participant achieved appropriate trust or not. For this reason, we also conducted a multilevel logistic regression per category, predicting proportions of the five trust categories separately. In our model, each round was treated as one observation, i.e., each row was one observation, with 15 rows per participant.

*Baseline Model*: We first created a baseline model, which comprised a random intercept per participant and the different explanation conditions. Next, we added the ''Wrong answers by the AI agent'' as additional fixed effects factor to our baseline model. Our dependent measure indicated whether this behavior is an appropriate trust behavior or not (similarly for other trust categories). Furthermore, we added a lag factor as a fixed effect to observe the effect of the previous round answer on the trust rating of the current round. The lag factor was coded as 1 if the previous trial was correct and 0 if not.

*Baseline Model plus Covariates*: We added three covariates ''Care about eating'' responses, ''Propensity to trust'' responses, and human confidence in estimating the calories (Q1) to our baseline model one-by-one. Since the $\chi^2$-based ANOVA comparison showed no significant improvement in the goodness-of-fit of the model upon adding the covariates and none of the covariates were significant predictors of any trust category, we decided not to include them in the models, see Table 4. For comparing the models for goodness-of-fit, AIC and BIC values are provided in the Appendix B, Table 11. We also report an marginal and conditional R-squared values, which indicates variance explained by both fixed and random effects; see Table 8.

*Appropriate Trust*: For the appropriate trust category, the "Fairness about risk" explanation was the only statistically significant predictor. The coefficient value of "Fairness" ($\beta = 0.591$, $p < .001$) is positive. Thus, we can say that when a participant interacted with the AI agent explaining with a focus on fairness through exposing risk and bias, the participant was more likely to achieve an appropriate level of trust in the AI Agent.

*Inconsistency*: For the inconsistency with a bad outcome trust category, we did not find any statistical significant predictor variable in our analysis. However, for the inconsistency with a good outcome trust category, the "Fairness about bias" explanation was again the only statistically significant predictor variable. The coefficient value of "Fairness about bias" ($\beta = -0.526$, $p < .001$) is negative. Thus, we can say that when participants interacted with the AI agent explaining with a focus on being fair by exposing bias and risk, the participants were less likely to end up in the inconsistency with a good outcome trust category.

*Under-trust and Over-trust*: For both the under-trust and over-trust categories, we did not find any statistically significant predictor variable in our analysis.

### 5.2 Effect of Different Principles of Integrity on Subjective Trust

In this subsection, we analyzed, how does human trust in the AI agent change given these different expressions of integrity principles (**RQ2**)? For this analysis, we performed a similar approach as RQ1 first to conduct descriptive statistics followed by inferential statistics where we focused on a multilevel regression model. Here, also, post-experiment questionnaire responses were analyzed to support the results and are reported in Section 6.2.

**Change in Trust Level Over Time:** We used a global trust meter to capture changes in trust over time. First, we calculated changes in human trust towards the AI agent over time by subtracting differences in trust meter values between every two subsequent rounds. As can be seen in Figure 5, trust in the AI agent dropped whenever the AI agent provided a wrong answer. We recorded an average drop of 15 points in trust score when a wrong answer was preceded by a right answer by the AI agent. This drop was more than twice the number of points when there were two wrong answers in a row, i.e., around 35 points. These results seem to confirm that the AI agent's accuracy influences trust.

**Predictors of Subjective Trust Scores:** Our dataset includes one row for each participant and one column for each variable or measurement on that participant. In the context of longitudinal data, this means that each measurement in time would have a separate row of its own, therefore,

Table 4. Results of GLMER Analysis for RQ1 (*: $p < .05$, **: $p < .01$, ***: $p < .001$)

| | Independent variables | Coefficient | | z value | Pr(>\|z\|) | Significance |
|---|---|---|---|---|---|---|
| | | $\beta$ | SE | | | |
| **Appropriate Trust** | | | | | | |
| (Intercept) | Participants | −0.262 | 0.135 | −1.931 | 0.053 | |
| | Fairness about bias | 0.591 | 0.162 | 3.650 | <0.001 | *** |
| | Honesty | 0.083 | 0.161 | 0.517 | 0.604 | |
| | Transparency | −0.009 | 0.162 | −0.594 | 0.552 | |
| | Wrong Answer | 0.077 | 0.097 | 0.793 | 0.427 | |
| | Lag (Wrong Answer) | −0.139 | 0.097 | −1.439 | 0.150 | |
| **Inconsistency (Bad outcome)** | | | | | | |
| (Intercept) | Participants | −2.572 | 0.240 | −10.692 | <0.001 | *** |
| | Fairness about bias | −0.395 | 0.274 | −1.441 | 0.150 | |
| | Honesty | −0.181 | 0.263 | −0.689 | 0.491 | |
| | Transparency | −0.170 | 0.263 | −0.645 | 0.519 | |
| | Wrong Answer | −0.291 | 0.196 | −1.486 | 0.137 | |
| | Lag (Wrong Answer) | 0.149 | 0.190 | 0.786 | 0.432 | |
| **Inconsistency (Good outcome)** | | | | | | |
| (Intercept) | Participants | −0.843 | 0.128 | −6.570 | <0.001 | *** |
| | Fairness about bias | −0.526 | 0.147 | −3.571 | <0.001 | *** |
| | Honesty | −0.225 | 0.142 | −1.583 | 0.113 | |
| | Transparency | −0.059 | 0.140 | −0.425 | 0.670 | |
| | Wrong Answer | −0.108 | 0.106 | −1.020 | 0.307 | |
| | Lag (Wrong Answer) | 0.162 | 0.106 | 1.519 | 0.128 | |
| **Under-trust** | | | | | | |
| (Intercept) | Participants | −2.180 | 0.224 | −9.720 | <0.001 | *** |
| | Fairness about bias | −0.485 | 0.289 | −1.681 | 0.092 | |
| | Honesty | 0.333 | 0.266 | 1.254 | 0.209 | |
| | Transparency | 0.246 | 0.268 | 0.915 | 0.360 | |
| | Wrong Answer | 0.236 | 0.141 | 1.667 | 0.095 | |
| | Lag (Wrong Answer) | −0.140 | 0.143 | −0.978 | 0.328 | |
| **Over-trust** | | | | | | |
| (Intercept) | Participants | −7.006 | 1.099 | −6.373 | <0.001 | ** |
| | Fairness about bias | 1.000 | 1.001 | 0.982 | 0.326 | |
| | Honesty | −0.125 | 1.014 | −0.124 | 0.901 | |
| | Transparency | 0.080 | 1.007 | 0.080 | 0.936 | |
| | Wrong Answer | −0.031 | 0.229 | −0.137 | 0.891 | |
| | Lag (Wrong Answer) | 0.184 | 0.233 | 0.790 | 0.430 | |

The marginal and conditional $R^2$ values are provided in the Appendix B for each model of the trust category.

we analyzed this data using a multilevel regression model following the instructions by Finch et al. [33, Chapter 5].

*Baseline Model*: We analyzed the global trust meter responses as our dependent variable to test the effect of different principles of integrity expressed through the explanations with a multilevel regression model with random intercept for trials. In addition, we added the current round correctness and lag as additional factors in our baseline model to test the effect of it on subjective trust

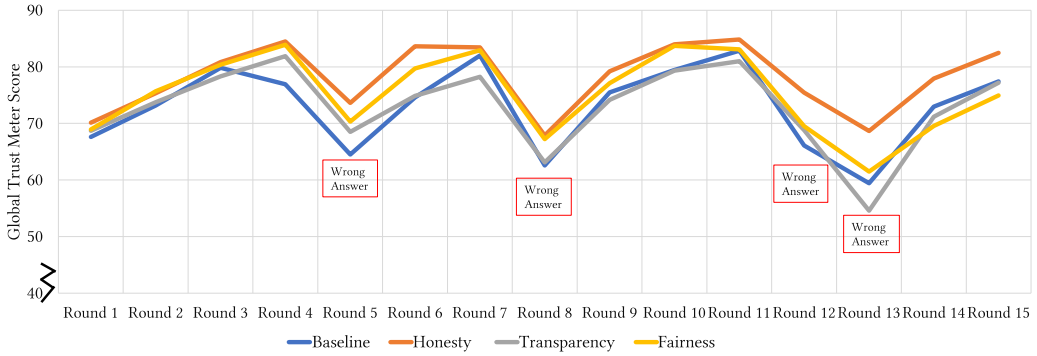


Fig. 5. Illustration of mean responses for changes in Global Trust Meter over 15 rounds. The red-colored boxes represent when the AI agent provided a wrong answer, i.e., rounds 5, 8, 12, and 13.

scores. Since our data is linear, we used the LMER method for this analysis with the lmerTEST package v3.1 [60].

*Baseline Model with a lag factor plus Covariates*: We now added fixed interaction effect between the correct/incorrect answer and the lag variable to the model. Furthermore, we also examined the two-way interaction effect between the correct/incorrect answer with different explanation types. This model was significantly better than the other two models for the goodness-of-fit, $Pr(> chisq) < 0.05$ (refer to Appendix B, Table 9 and 10, for further details). Hence, we finalized this model as reported in Table 5.

Following the same procedure as RQ1, we further explored adding same covariates to our model. Adding these covariates did not improve our model and, therefore, we did not include those variables in our final model. Finally, we added human comfort and usefulness of explanations ratings to the model and found that only the usefulness of explanations helps in improving our model.

Based on the regression results, we can observe that the honesty explanation is a significant predictor of the trust score compared to other explanations expressing integrity ($\beta = 7.84$, $p < .05$), i.e., participants who saw the honesty explanation rated their subjective trust in the AI agent higher than the other conditions. Furthermore, as shown in Table 5, both the correct/incorrect answer and the lag variable are statistically significant predictors of the subjective trust ratings ($p < .05$). This result confirms our intuition observed from Figure 5, where the effect of the correct/incorrect answer on the trust scores can be observed. Interestingly, the significance of the lag variable shows the effect of the previous round correctness on the current round trust score. In other words, as it is important to study the effect of the correct answer on the trust score for the current score, it is equally important to study how the AI agent performed a round before to observe the changes in the trust score.

Additionally, our results show that the interaction effect between the correct/incorrect answer and the lag variable is significant ($\beta = -3.38$, $p < .05$). Given that the sign on the interaction coefficient is negative, we would conclude that there is a buffering or inhibitory effect. Analyzing the correct/incorrect answer, lag, and their interaction reveals the drop and restoration of global trust ratings. For instance, two consecutive correct trials yield a combined score of 21.44, while a correct trial followed by an incorrect one results in a high initial drop of 7.77. Similarly, an incorrect trial followed by a correct one leads to a recovery to 17.05, almost reaching 21.44 again. Two consecutive incorrect trials cause a complete drop to 0, followed by a gradual recovery to 7.77, 17.05, and 21.44. These findings align with the results in Figure 5.

Moreover, there is a significant interaction effect between the correct/incorrect answer and the honesty explanation ($\beta = -4.63$, $p < .05$). This indicates that the impact of errors is smaller in the

Table 5. Results of LMER Analysis for RQ2 (*: $p < .05$, **: $p < .01$, ***: $p < .001$) with LMERTest R Package

| | Independent variables | Coefficient | | t value | Pr(>\|t\|) | Significance |
|---|---|---|---|---|---|---|
| | | $\beta$ | SE | | | |
| **Global Trust Ratings** | | | | | | |
| (Intercept) | | 46.11 | 2.90 | 15.88 | <0.001 | *** |
| Participants | Round | 0.01 | 0.08 | 0.12 | 0.907 | |
| | Fairness about bias | 4.06 | 3.20 | 1.27 | 0.205 | |
| | Honesty | 7.84 | 3.20 | 2.45 | 0.015 | * |
| | Transparency | 0.68 | 3.20 | 0.21 | 0.831 | |
| | Correct/Incorrect Answer | 17.05 | 1.68 | 10.14 | <0.001 | *** |
| | Lag Correct/Incorrect | 7.77 | 1.30 | 6.00 | <0.001 | *** |
| | Correct/Incorrect*Lag | − 3.38 | 1.46 | − 2.31 | 0.021 | * |
| | Correct/Incorrect*Fairness | − 2.71 | 1.80 | − 1.51 | 0.132 | |
| | Correct/Incorrect*Honesty | − 4.63 | 1.80 | − 2.57 | 0.010 | * |
| | Correct/Incorrect*Transparency | − 1.55 | 1.80 | − 0.86 | 0.391 | |
| | Usefulness of Explanations | 1.72 | 0.18 | 9.45 | <0.001 | *** |
| | Marginal $R^2$ | | | | 0.136 | |
| | Conditional $R^2$ | | | | 0.534 | |

honesty condition, as depicted in Figure 5. Also, usefulness of explanations is a predictor of global trust ratings ($\beta = 9.45$, $p < .0001$). This result means the participants found the explanations helpful in adjusting their trust levels after each round.

### 5.3 Effect of Different Principles of Integrity on Human's Decision-making and Usefulness of Explanations

In this subsection, we analyzed, how do different expressions of integrity principles influence the human's decision-making, and do people feel these explanations are useful in making a decision (**RQ3)**? For this analysis, we performed a similar approach as in RQ2.

**Descriptive statistics:** We used human comfort ratings (Q4) and usefulness of explanations ratings (Q5) by participants to analyze our responses for RQ3. These ratings were measured after each trial. Therefore, we followed the same analysis method as for RQ2. For the human comfort ratings, we did not find any major differences among the four conditions; refer to Figure 7, Appendix A. The mean ratings for the baseline condition was 6.178 (1.981), for honesty 6.285 (1.863), for transparency 6.246 (1.811), and for fairness 6.128 (1.948). Similarly, for the helpfulness of explanations ratings, we also did not find any major differences among the four conditions; refer to Figure 8, Appendix A. The mean ratings for the baseline condition was 6.333 (2.053), for honesty 6.675 (1.764), for transparency 6.486 (1.845), and for fairness 6.423 (1.831).

**Predictors of Comfort and Explanations Helpfulness:** We analyzed the human comfort ratings and usefulness of explanations responses as our dependent variable to test the effect of different principles expressed through the explanations with a multilevel regression model with random intercept for participants. We followed the similar model as for RQ2 in analyzing the results of this RQ. Adding the covariates from RQ1 did not improve both our models (human comfort and explanations help). Also, adding the interactions as in Table 5 was not helpful in improving the model statistics. Therefore, we did not include them in our final models. We report the regression model of predicting the usefulness of explanations in Table 6 and human comfort in Table 7, Appendix B.

Table 6. Results of LMER Analysis for RQ3—Helpfulness of Explanations (*: $p < .05$, **: $p < .01$, ***: $p < .001$)

| | Independent variables | Coefficient | | t value | Pr(>\|t\|) | Significance |
|---|---|---|---|---|---|---|
| | | $\beta$ | SE | | | |
| **Explanations Help** | | | | | | |
| (Intercept) | | 2.72 | 0.25 | 10.88 | <0.001 | *** |
| Participants | Round | 0.01 | 0.01 | 1.24 | 0.217 | |
| | Fairness about bias | 0.09 | 0.19 | 0.45 | 0.654 | |
| | Honesty | 0.29 | 0.19 | 1.48 | 0.141 | |
| | Transparency | 0.13 | 0.19 | 0.69 | 0.491 | |
| | Correct/Incorrect Answer | − 0.05 | 0.08 | − 0.60 | 0.548 | |
| | Lag Correct/Incorrect | − 0.14 | 0.08 | − 1.87 | 0.062 | |
| | Trust Score | 0.02 | 0.00 | 10.30 | <0.001 | *** |
| | Human Comfort | 0.34 | 0.02 | 17.93 | <0.001 | *** |
| | Marginal $R^2$ | | | | 0.230 | |
| | Conditional $R^2$ | | | | 0.391 | |

Based on Table 6, we can observe that the trust score is a significant predictor of the usefulness of the explanations ($\beta = 0.02$, $p < .001$), i.e., participants who rated their subjective trust in the AI agent could have found the explanations provided by it more helpful. Similarly, we found that human comfort in decision-making is another significant predictor of the usefulness of the explanations score ($\beta = 0.34$, $p < .001$). None of the other covariates were found to be significant predictors of the human comfort score except the helpfulness of explanations ($\beta = 0.35$, $p < .001$).

## 6 DISCUSSION

Our results offer three major contributions for discussion in the field of the Human-AI interaction:

(1) We can measure appropriate trust through a formal computation method in the context of a specific task.
(2) Appropriate trust can be enhanced by providing expressions of fairness principle of integrity in the context of human-AI interaction. Furthermore, appropriate trust builds up over time and recovers slowly if an AI agent provides an incorrect output.
(3) Subjective trust builds up and recovers better by providing expressions of honesty in human-AI interaction.

Therefore, in this section, we will discuss our results about how the explanations expressing different integrity principles influenced appropriate trust. Next, we will discuss how participants perceived the AI agent's advice and made their decision based on theories from psychology and social sciences, which possibly led them to select the AI agent. Finally, we will discuss the limitations of our work and possible future directions.

### 6.1 Expressions of Integrity and Appropriate Trust

We found that the "Fairness about bias" explanations were the most effective for fostering appropriate trust in the AI agent. We know from previous work by Asan et al. [4] that knowing about biases can influence human trust, which perhaps also explains why trust becomes more appropriate if human can intervene in AI decision-making.

A closer look at our findings shows that, in our case, the explanations highlighting potential bias and risks actually improved appropriate trust through increasing trust rather than decreasing

it. This makes intuitive sense, as fairness explanations could have triggered more cognitive effort resulting in increase of people's cognitive thinking for engaging analytically with the explanation [13]. Furthermore, recent education research has shown that students' actual learning and performance was better with the more cognitively demanding instructions [24]. Overall, our findings seem to support the proposition that we should be building explainable and bias-aware AI systems to facilitate rapid trust calibration, leading to building appropriate trust in human-AI interaction [101].

Interestingly, irrespective of which integrity principle was highlighted, explanations seem to have helped our participants in correcting under-trust and over-trust (see Figure 3). In particular, being explicit about potential biases and risks actually decreased inconsistent behavior with a good outcome over the other explanation types in some cases (including those cases where trust was appropriate). A possible reason is that these explanations exposed potential bias(es) in the data or the model, which could have convinced the participants to follow the AI agent. For example, P62 reported that *"If the AI Assistant says dataset is biased, then [it's] true I suppose and it's more trustworthy than my common sense because I haven't seen the data, so I will stick to my initial trust decision" (P62, Fairness about bias condition).* Similarly, P133 reported *"I feel like the results of [the model] were strange hence I went with my decision first but I was wrong, so next time for a similar round I choose the [AI] Assistant and it was right. Hence, I decided to follow him [AI Agent]!" (P133, Fairness about bias condition).*

Another finding of our study was that irrespective of what principle of integrity was expressed in the explanation, around 30% of the time participants ended up in the inconsistency (good and bad outcome) trust category. This shows that even when participants reported that they trusted the AI agent to be better than themselves, they still quite often chose not to rely on it. Based on our exit interviews, we found that participants acted inconsistently several times during the experiment to increase their score leading to winning the gift voucher. For example, P20 told us "*I think [AI Agent] it is better in identifying this dish, but it was also wrong with a similar dish in one of the previous rounds, so I will choose myself because I do not want to lose any points.*" Similarly, P77 said "*Ahh, I was just checking if I say I trust [AI Agent] him but do not go with him then what will happen. If it turns out to be good, I will do this again to keep my score up.*"

We found none of the covariates "Care about eating" and "propensity to trust" a predictor of subjective trust score and any trust category. For "Care about eating," a potential reason could be that people who rated higher on caring about their eating behavior were more aware of the different ingredients with their calories level that were known to them and vice versa. Given the images of the food items in our experiment were diverse, this could have impacted their skills to judge the calories well. For example, P97 with a score of 10 for the "Care about eating" question reported that "*I am very picky about what I eat as I need my balanced diet. However, this task is not easy as it has many international cuisines!*" For, "propensity to trust," one possible explanation can be that this dispositional covariate became less important as system experience increased. Alternatively, this covariate could influence trusting behaviors more than trusting beliefs. More research is needed on the effect of propensity to trust factors over time.

### 6.2 Subjective Trust, Helpfulness, and Comfort

Subjective trust is not the same as appropriate trust [118]. Chen et al. [18] identified in their study that participant's objective trust calibration (proper uses and correct rejections) improved as intelligent agent became more transparent. However, their subjective trust did not significantly increase. The "Fairness about bias" explanation in our work helped in fostering appropriate trust in the AI agent. However, it did not necessarily improve participant's (subjective) trust. This result

is in line with Buçinca et al. [13], who showed that there exists a tradeoff between subjective trust and preference in a system of human+AI decision-making.

From Figure 5 and Table 5, it is evident that the subjective trust ratings for the "Honesty" explanations are significantly higher compared to the other explanation types. This observation can be explained by the explicit references to honesty by the AI agent as reported by P102, *"It [AI Agent] mostly talks about being honest and based on all rounds—I think it is, so I trust it" (P102, Honesty condition).* We can recall that the AI agent in the "Honesty" condition expressed its honesty by stating it cared about honesty and adding further information about uncertainty in the decision-making. This expression of honesty resonates with Wilson [115], who argued that as long as communication is performed in an honest way, it produces ecological integrity affecting trust.

We also found the effect of the current and the previous round correctness on the subjective trust ratings; refer to Table 5. This result is echoed from a prior study by Tolmeijer et al. [99], who showed that system accuracy influences the trust development in an AI system. Furthermore, the effect of the previous round correctness, i.e., the lag in Table 5, had an influence on the trust score as well. This result indicates that trust is not only influenced by how the system is performing now but also on how it performed before. Human trust develops over time and depends on many factors. Also, each interaction with a system can alter the trust in that system. For example, Holliday et al. [41] looked at trust formation within one user session; they found that the impression of system reliability at each time point shapes trust. Our results align with van't Wout et al. [105], who show that the outcome of a previous round (whether the trust was repaid or abused) affected how much a participant trusts another participant to send money.

Turning to the transparency explanations, based on post questionnaire responses, the participants found the visual part of the explanation difficult to follow. For example, *"I can see there is best, good and unsure match but I have no idea it really helps as everything looks almost same!" (P140, Transparency condition).* Additionally, we believe that the combination of visual with textual explanations may have hampered understandability as reported by P17 *"That's simply too much of information for me!" (P17, Transparency condition).*

Overall, trust scores exhibit a consistent level of stability, particularly an initial overall level of trust that remains steady over time, except in cases where an error occurs (Figure 5). This is in line with our intuition of how trust works. Interestingly, while an increase of trust between rounds three and four was expected, trust recovers to same levels between rounds six and seven and nine and ten. A potential explanation can be that the AI agent's early impressions positively influenced the AI agent's perceived reliability, leading to increased trust even after inaccurate advice.

The result in Table 6 demonstrates no effect of type of explanations on participant's usefulness of explanations ratings. However, we found that participant's trust and human comfort scores significantly predicted the usefulness of explanations ratings. We can understand this result as if an explanation was helpful; participants often rated their trust and comfort in the decision-making process higher than the non-helpful explanations.

We also found that participant's decision-making comfort levels were similar across conditions. However, the explanations score significantly predicted the participant's comfort level. A potential reason might be that other individual factors more strongly influence the subjective notion of the comfort of decision-making than the differences between our explanations. Another possible explanation is that different types of explanations by the AI agent did not necessarily improve the comfort level but only assisted in decision-making. A previous study focusing on integrity among colleagues reported that showing integrity did not increase the comfort level of employees to rely on each other [113]. This result aligns with our findings, where it is hard to establish human comfort by expressing principles related to integrity.

### 6.3 Understanding Human Psychology for AI Assistant's Advice Utilization

Advice utilization has been studied in the literature of psychology to understand how humans utilize the advice given by others [70]. Jodlbauer and Jonas [51] found that while three different dimensions of trust (competence, benevolence, and integrity) mediate between advisor and listener, for the acceptance of advice, trust in advisor integrity played the strongest mediating role in human-human interaction.

Given that all the AI agents in our user study had the same competence level, the only difference was what principle of integrity was highlighted in the explanation of the AI agent. This difference partly explains why integrity expressions of fairness through exposing potential bias and risk help understand appropriate trust in our study. Furthermore, this difference partly explains how integrity expressions of honesty about uncertainty in decision-making help understand users' subjective trust in our study.

The theory by Bazerman and Moore [8] can help us partly understand why explanations exposing potential bias and risk were significantly different from the other explanations used in this study. They showed that humans are limited in their rationality and are often subject to cognitive bias. Furthermore, when decisions involve risks based on unbiased advice and people cannot weigh all relevant information, decision-makers often use the advice [8] that helps in reducing their own bias. Therefore, participants' trust in the "fairness about risk" condition was more appropriate compared to other conditions. For example, P73 reported, *"I was not sure about different type of vegetables in the salad but the AI told me correctly that it was also not sure, hence I decided not to trust it and went with my best possible option—which was eventually correct!"*.

### 6.4 Reflections on Design Considerations for Building Appropriate Trust

In the prior research, appropriate trust is often linked with [not] relying on the AI system when it makes a [in]correct decision. This notion of appropriate trust heavily relies on the capability of the AI system leaving out other factors that can influence trust, such as integrity or benevolence. Here, our work serves as an example of how expressing different principles related to integrity through explanations can establish appropriate trust in human-AI interaction. Therefore, an essential focus of designing AI for fostering appropriate trust should be both on the capability as well as the integrity of the AI system. However, this comes with the challenge of obtaining accurate measurement information regarding the machine learning models' performance, bias, fairness, and inclusion considerations.

Lord Kelvin has promoted measurement with his memorable statement: "If you cannot measure it, you cannot improve it" [54]. There is much discussion on the AI systems to be appropriately trusted. However, there are very few suggestions for measuring the appropriate trust. Part of this lack of literature on measurement is because trust is subjective in nature. What seems an appropriate trust for person A will not be appropriate for person B. Nevertheless, it is also crucial for humans to calibrate their trust, recognizing that AI systems can never be 100% trustworthy. Likewise, we made an attempt to capture trust into various categories (appropriate, over-/under-trust, inconsistency) through formal definitions.

We believe that our proposed formal definitions can help facilitate communication between researchers, practitioners, and stakeholders by providing a common language and understanding of what is meant by measuring appropriate trust. Furthermore, it can set clear expectations for how trust should be measured, can promote a better understanding of what trust means, and what aspects of trust should be considered [10]. We hope this work highlights the need for guidelines to incorporate a method to capture appropriate trust and develop an understanding of human decision-making with psychological theories such as advice utilization.

### 6.5 Limitations and Future Work

Our work limits itself to exclusive decision-making, which does not represent the full spectrum of possible human-AI interaction. Our task was inspired by scenarios in which a human needs to make a conscious choice to either follow the system advice or their own; such as the autopilot mode or cruise control in a car. Therefore, our findings may not generalize to every scenario such as human-AI teaming, where the focus is more on the collaboration. Additionally, in our definition of appropriate trust, we did not further explore the reasons for the selections made by the human. Interesting notions for further study are how our notions of appropriate trust can be influenced by the delegation of responsibility, focusing on different choices people make in the delegation. For example, people are more likely to delegate a choice when it affects someone so as not to take the blame if something goes wrong [95].

In our user study, we used images of various food items for estimation of food calories based on a machine learning model. In our day-to-day situations, people hardly use such technological advances. Therefore, the level of realism can be further improved in future studies. Furthermore, our users got 15 trials in the same condition, which could have led to possible learning or fatigue effects even though we provided a break after seven rounds. Also, the order of the wrong AI advice was same across the conditions, which made it hard for us to control the possible fatigue effects.

We have utilized situation vignettes to craft our explanations. In our work, custom build explanations to highlight different principles related to integrity were better suited to our user study, i.e., by explicitly revealing the importance of individual notions of integrity (honesty, transparency, and fairness) in a calories estimation task. In this, we attempted to keep other variables (e.g., length) mostly the same, but, for instance, it was inevitable that the baseline explanation would be shorter. The style was controlled for in some way by having the same authors for all explanations, but here, too, differences might exist between conditions. For instance, the "fairness about risk" explanation might have been a little more technical, as it explained where in the process risks could come from (e.g., bias in training data). Although we cannot exclude such influences, we would argue that such slight differences will always be inevitable when expressing different principles in explanations. More research on, e.g., style of writing, length, would be relevant to further control for such factors [106].

Finally, our explanations express the related principles of integrity in one specific way, and different methods of expressing these might have different effects on trust than what we found. However, with this work, we show a method for the AI agent to express its integrity in the form of explanations, and our aim for this research was not to design effective explanations but to study how different expressions of integrity can help in building appropriate trust.

A future research direction to scale this work could look at how we can create vignettes by systematically combining actions of the agent based on the affect control theory [38] in real time. For example, one could adopt ensemble machine learning methods, as they are shown to perform well and generalize better for generating action-based situations [26]. One could also look at PsychSim [89] framework, which combines two established agent technologies: decision-theoretic planning and recursive modeling for crafting explanations using machine learning models.

Furthermore, the understandability of explanations might be further enhanced by design specialists and tested by crowdsourcing with a diverse demographic sampling. Broader findings would further enable designers to craft explanations to make AI systems more understandable and trustworthy. Finally, further work can explore trusting behavior targeting both integrity and benevolence as antecedents of trust.

## 7 CONCLUSION

Our user study was a means to employ the formal definition of appropriate trust and understand how expressions of principles related to integrity through explanations can help in fostering appropriate trust. In this article, we (a) provided a formal definition of appropriate trust following the interpersonal perspective of trust, (b) investigated different ways of expressing principles related to integrity through explanations—honesty about uncertainty; transparency about the decision-making process; and fairness in terms of being open about potential bias and risk by an AI agent, and (c) showed the effect of these different types of integrity-based explanations on the end-user's appropriate trust. Our task involved an exclusive decision-making process where participants were required to select either themselves or rely on the AI agent for the task. Our results show a strong correlation between expressing integrity focused on fairness in openness about biases and appropriate trust. In summary, the two key takeaway messages of this work are (1) a measurement method for appropriate trust in exclusive decision-making task and (2) expressing integrity principles in explanations given by an AI agent has the potential to improve end-users' appropriate trust and enhance the appropriate use of AI systems.

## APPENDICES

## A APPENDIX


Fig. 6. A correlation plot between trust categories and integrity conditions. Positive residuals are in blue and specify an attraction (positive association). Negative residuals are in red, implying a repulsion (negative association). The relative contribution of each cell to the total chi-square score provides an indication of the nature of the dependency between trust categories and conditions.

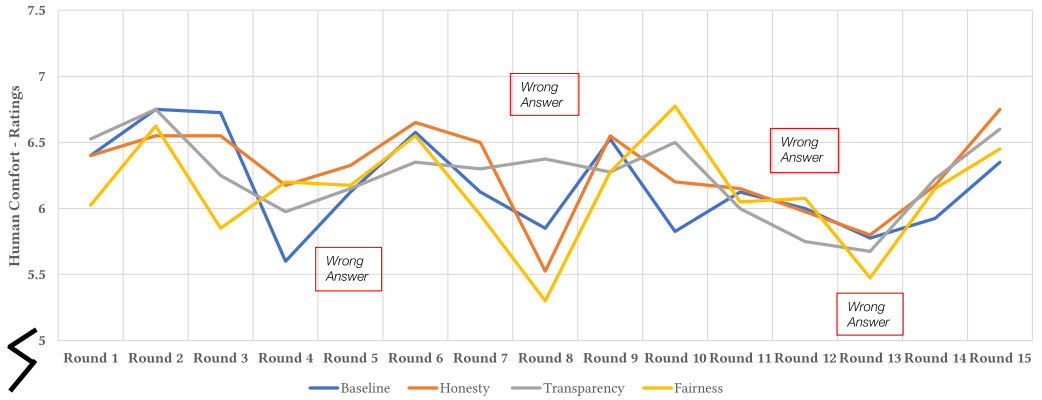


Fig. 7. Illustration of mean responses for changes in human comfort in decision-making ratings over 15 rounds. The red-colored boxes represent when the AI agent provided a wrong answer, i.e., rounds 5, 8, 12, and 13.

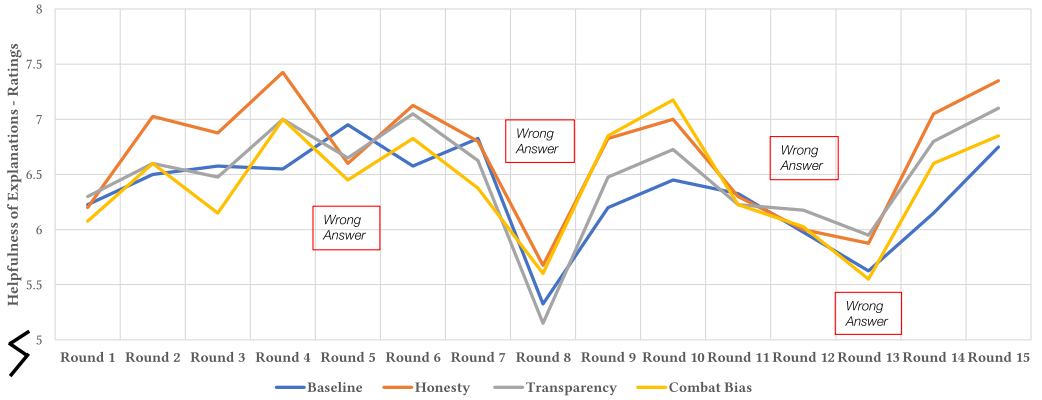


Fig. 8. Illustration of mean responses for changes in helpfulness of explanations ratings over 15 rounds. The red-colored boxes represent when the AI agent provided a wrong answer, i.e., rounds 5, 8, 12, and 13.

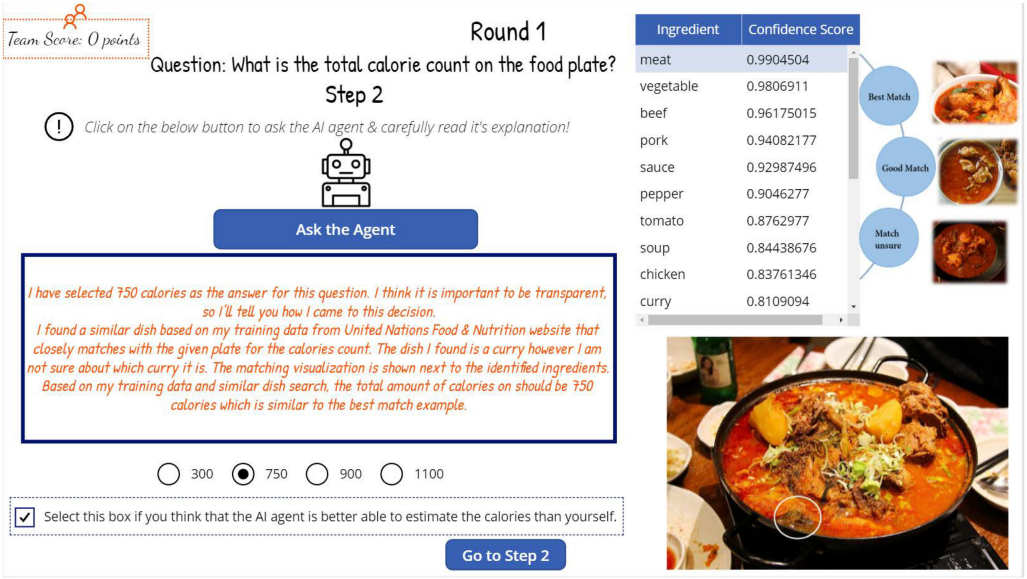


Fig. 9. Screenshot of transparency condition of the user study. This condition provided visualization of confidence scores in terms of best, good, and an unsure match (refer to top right corner).

## B APPENDIX

Table 7. Results of LMER Analysis for RQ3—Helpfulness of Explanations (*: $p < .05$, **: $p < .01$, ***: $p < .001$)

| | Independent variables | Coefficient | | t value | Pr(>\|t\|) | Significance |
|---|---|---|---|---|---|---|
| | | $\beta$ | SE | | | |
| **Human Comfort** | | | | | | |
| (Intercept) | | 4.08 | 0.27 | 15.31 | <0.001 | *** |
| Participants | Round | 0.01 | 0.01 | 0.26 | 0.792 | |
| | Fairness about bias | − 0.14 | 0.23 | − 0.58 | 0.562 | |
| | Honesty | − 0.04 | 0.24 | − 0.16 | 0.870 | |
| | Transparency | 0.00 | 0.23 | 0.02 | 0.985 | |
| | Correct/Incorrect Answer | − 0.16 | 0.08 | − 2.08 | 0.561 | |
| | Lag Correct/Incorrect | − 0.28 | 0.08 | − 1.71 | 0.069 | |
| | Trust Score | 0.01 | 0.00 | 1.40 | 0.162 | |
| | Explanation Help | 0.35 | 0.02 | 17.36 | <0.001 | *** |
| | Marginal $R^2$ | | | | 0.150 | |
| | Conditional $R^2$ | | | | 0.394 | |

Table 8. Marginal and Conditional $R^2$ Values for Regression Model of RQ1

| Model | Marginal $R^2$ | Conditional $R^2$ |
|---|---|---|
| Appropriate Trust | 0.021 | 0.082 |
| Inconsistency (Bad outcome) | 0.011 | 0.098 |
| Inconsistency (Good outcome) | 0.014 | 0.032 |
| Under-trust | 0.028 | 0.199 |
| Over-trust | 0.010 | 0.848 |

Table 9. AIC and BIC Statistics for the Regression Models of RQ2

| Model | AIC | BIC |
|---|---|---|
| Baseline | 18,617 | 18,662 |
| Baseline+Lag | 18,566 | 18,618 |
| Baseline+Lag+Interactions | 18,565 | 18,639 |
| Baseline+Lag+Interactions+Helpfulness of Explanations | 18,481 | 18,561 |

Table 10. Regression Models Comparisons of RQ2

| Model | Baseline | | Baseline+Lag | |
|---|---|---|---|---|
| | Chi Square | Pr(>Chisq) | Chi Square | Pr(>Chisq) |
| Baseline+Lag+Interactions | 62.032 | <0.001 | 9.701 | 0.045 |

Table 11. AIC and BIC Statistics for the Regression Models of RQ1

| Model | AIC | BIC |
|---|---|---|
| **Appropriate Trust** | | |
| Baseline (Correct/Incorrect Answer+Lag) | 3,037.5 | 3,077.5 |
| Baseline+Covariate 1 (Care about eating) | 3,039 | 3,084.7 |
| Baseline+Covariate 2 (Propensity to Trust) | 3,039.3 | 3,085 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 3,039.1 | 3,084.8 |
| Baseline+Covariate 4 (Human Comfort) | 3,039.3 | 3,085 |
| **Inconsistency with a bad outcome** | | |
| Baseline (Correct/Incorrect Answer+Lag) | 1,140.7 | 1,180.7 |
| Baseline+Covariate 1 (Care about eating) | 1,142.6 | 1,188.3 |
| Baseline+Covariate 2 (Propensity to Trust) | 1,142.5 | 1,188.2 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 1,139.9 | 1,185.6 |
| Baseline+Covariate 4 (Human Comfort) | 1,142.7 | 1,188.4 |
| **Inconsistency with a good outcome** | | |
| Baseline (Correct/Incorrect Answer+Lag) | 2,653 | 2,693 |
| Baseline+Covariate 1 (Care about eating) | 2,651.7 | 2,697.4 |
| Baseline+Covariate 2 (Propensity to Trust) | 2,653.6 | 2,699.3 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 2,654.9 | 2,700.6 |
| Baseline+Covariate 4 (Human Comfort) | 2,654 | 2,699.7 |
| **Under-trust** | | |
| Baseline (Correct/Incorrect Answer+Lag) | 1,671.1 | 1,711.1 |
| Baseline+Covariate 1 (Care about eating) | 1,672 | 1,717.7 |
| Baseline+Covariate 2 (Propensity to Trust) | 1,673.1 | 1,718.8 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 1,671.5 | 1,717.2 |
| Baseline+Covariate 4 (Human Comfort) | 1,670.3 | 1,716.1 |
| **Over-trust** | | |
| Baseline (Correct/Incorrect Answer+Lag) | 822.8 | 862.8 |
| Baseline+Covariate 1 (Care about eating) | 822 | 867.7 |
| Baseline+Covariate 2 (Propensity to Trust) | 824.4 | 870.1 |
| Baseline+Covariate 3 (Usefulness of Explanations) | 824.6 | 870.3 |
| Baseline+Covariate 4 (Human Comfort) | 824.2 | 869.9 |

AIC is best for prediction, as it is asymptotically equivalent to cross-validation. BIC is best for explanation, as it is allows consistent estimation of the underlying data-generating process.

## ACKNOWLEDGMENTS

We thank Luciano Siebert, Pei-Yu Chen, Nele Albers, Enrico Liscio, Tim Draws, Ruben Verhagen, and anonymous reviewers for their contribution in iterations of the project.

## REFERENCES

[1] 2017. *Ethically Aligned Design—A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2, 2017.* The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[2] Google PAIR. 2019. People + AI Guidebook. Retrieved May 18, 2021 from https://pair.withgoogle.com/guidebook/

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*, Association for Computing Machinery, Glasgow, Scotland Uk, 1–13. DOI: https://doi.org/10.1145/3290605.3300233

[4] Onur Asan, Alparslan Emrah Bayrak, Avishek Choudhury. 2020. Artificial intelligence and human trust in healthcare: Focus on clinicians. *J. Med. Internet Res.* 22, 6 (2020), e15154. https://www.jmir.org/2020/6/e15154

[5] Giselle A. Auger. 2014. Trust me, trust me not: An experimental analysis of the effect of transparency on organizations. *J. Pub. Relat. Res.* 26, 4 (2014), 325–343.

[6] Alan D. Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *J. Verb. Learn. Verb. Behav.* 14, 6 (1975), 575–589.

[7] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[8] Max H. Bazerman and Don A. Moore. 2012. *Judgment in Managerial Decision Making*. John Wiley & Sons.

[9] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *J. Assoc. Inf. Syst.* 6, 3 (2005), 4.

[10] Rajeev Bhattacharya, Timothy M. Devinney, and Madan M. Pillutla. 1998. A formal model of trust based on outcomes. *Acad. Manag. Rev.* 23, 3 (1998), 459–472.

[11] Jens Blechert, Adrian Meule, Niko A. Busch, and Kathrin Ohla. 2014. Food-pics: An image database for experimental research on eating and appetite. *Front. Psychol.* 5 (2014), 617.

[12] Tibor Bosse, Catholijn M. Jonker, Jan Treur, and Dmytro Tykhonov. 2007. Formal analysis of trust dynamics in human and software agent experiments. In *Proceedings of the International Workshop on Cooperative Information Agents*. Springer, 343–359.

[13] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (Apr.2021), 21 pages. DOI: https://doi.org/10.1145/3449287

[14] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the International Conference on Healthcare Informatics*. IEEE, 160–169.

[15] John K. Butler Jr. 1991. Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *J. Manag.* 17, 3 (1991), 643–663.

[16] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.

[17] Cristiano Castelfranchi and Rino Falcone. 2016. Trust & self-organising socio-technical systems. In *Trustworthy Open Self-organising Systems*. Springer, 209–229.

[18] Jessie Y. C. Chen and Michael J. Barnes. 2014. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Trans. Hum.-mach. Syst.* 44, 1 (2014), 13–29.

[19] Andy Cockburn and Carl Gutwin. 2009. A predictive model of human performance with scrolling and hierarchical lists. *Hum.–comput. Interact.* 24, 3 (2009), 273–314.

[20] Marvin S. Cohen, Raja Parasuraman, and Jared T. Freeman. 1998. Trust in decision aids: A model and its training implications. In *Proceedings of the Command and Control Research and Technology Symposium*.

[21] Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert A. Lafrance, Nick Cramer, Kristin Cook, and Samuel Payne. 2016. Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE Trans. Visualiz. Comput. Graph.* 23, 1 (2016), 271–280.

[22] Ewart de Visser and Raja Parasuraman. 2011. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *J. Cogn. Eng. Decis. Mak.* 5, 2 (2011), 209–231.

[23] 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. *Int. J. Soc. Robot.* 12, 2 (2020), 459–478.

[24] Louis Deslauriers, Logan S. McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. 2019. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proc. Nat. Acad. Sci.* 116, 39 (2019), 19251–19257.

[25] S. Kate Devitt. 2018. Trustworthiness of autonomous systems. In *Foundations of Trusted Autonomy*. Springer, Cham, 161–184.

[26] Thomas G. Dietterich. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.* 40, 2 (2000), 139–157.

[27] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. https://www.sciencedirect.com/science/article/pii/S1071581903000387

[28] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[29] Fredrick Ekman, Mikael Johansson, and Jana Sochor. 2017. Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Trans. Hum.-mach. Syst.* 48, 1 (2017), 95–101.
[30] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Meth.* 41, 4 (2009), 1149–1160.
[31] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Sci. Eng. Ethics* 26, 6 (2020), 3333–3361.
[32] Carolina Ferreira Gomes Centeio Jorge, Siddharth Mehrotra, Myrthe L. Tielman, and Catholijn M. Jonker. 2021. Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. In *Proceedings of the 22nd International Workshop on Trust in Agent Societies*.
[33] W. Holmes Finch, Jocelyn E. Bolin, and Ken Kelley. 2019. *Multilevel Modeling Using R*. CRC Press.
[34] Michael W. Floyd, Michael Drinkwater, and David W. Aha. 2014. How much do you trust me? Learning a case-based model of inverse trust. In *Proceedings of the International Conference on Case-based Reasoning*. Springer, 125–139.
[35] M. Lance Frazier, Paul D. Johnson, and Stav Fainshmidt. 2013. Development and validation of a propensity to trust scale. *J. Trust Res.* 3, 2 (2013), 76–97.
[36] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *Proceedings of the International Symposium on Collaborative Technologies and Systems*. IEEE, 106–114.
[37] Anthony Giddens. 2013. *The Consequences of Modernity*. John Wiley & Sons.
[38] David R. Heise. 1979. *Understanding Events: Affect and the Construction of Social Action*. Cambridge University Press New York.
[39] Robert R. Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive Systems Engineering* (1st Edition), CRC Press (2017), 28 pages.
[40] Heike Hofmann. 2000. Exploring categorical data: Interactive mosaic plots. *Metrika* 51, 1 (2000), 11–26.
[41] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 164–168.
[42] Linda Childers Hon and James E. Grunig. 1999. Guidelines for measuring relationships in public relations. (1999). https://instituteforpr.org//wp-content/uploads/Guidelines_Measuring_Relationships.pdf
[43] Chong Sun Hong and Tae Gyu Oh. 2021. Correlation plot for a contingency table. *Commun. Stat. Applic. Meth.* 28, 3 (2021), 295–305.
[44] Marc W. Howard and Michael J. Kahana. 1999. Contextual variability and serial position effects in free recall. *J. Experim. Psychol.: Learn., Mem. Cogn.* 25, 4 (1999), 923.
[45] Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. 2018. Establishing appropriate trust via critical states. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)*. IEEE, 3929–3936.
[46] Leo W. J. C. Huberts. 2018. Integrity: What it is and why it is important. *Pub. Integ.* 20, sup1 (2018), S18–S32.
[47] Aya Hussein, Sondoss Elsawah, and Hussein A. Abbass. 2020. Trust mediating reliability–reliance relationship in supervisory control of human–swarm interactions. *Hum. Fact.* 62, 8 (2020), 1237–1248.
[48] Brett W. Israelsen and Nisar R. Ahmed. 2019. "Dave... I can assure you... that it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Comput. Surv.* 51, 6 (2019), 1–37.
[49] T. H. Jeavons. 2001. *Ethics in Nonprofit Management*. Routledge, 108–119.
[50] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md Abdullah Al Fahim. 2018. Initial trustworthiness perceptions of a drone system based on performance and process information. In *Proceedings of the 6th International Conference on Human-agent Interaction*, 229–237.
[51] Barbara Jodlbauer and Eva Jonas. 2011. Forecasting clients' reactions: How does the perception of strategic behavior influence the acceptance of advice? *Int. J. Forecast.* 27, 1 (2011), 121–133.
[52] Devon Johnson and Kent Grayson. 2005. Cognitive and affective trust in service relationships. *J. Bus. Res.* 58, 4 (2005), 500–507.
[53] Wiard Jorritsma, Fokie Cnossen, and Peter M. A. van Ooijen. 2015. Improving the radiologist–CAD interaction: Designing for appropriate trust. *Clin. Radiol.* 70, 2 (2015), 115–122.
[54] Lord Kelvin. 1883. William Thomson. *Electrical Units of Measurement Popular Lectures And Adresses* 1 (1883).
[55] Mohammad T. Khasawneh, Shannon R. Bowling, Xiaochun Jiang, Anand K. Gramopadhye, and Brian J. Melloy. 2003. A model for predicting human trust in automated systems. *Origins* 5 (2003).
[56] Sara Kiesler and Jennifer Goetz. 2002. Mental models of robotic assistants. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. 576–577.

[57] Wonjoon Kim, Nayoung Kim, Joseph B. Lyons, and Chang S. Nam. 2020. Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Appl. Ergon.* 85 (2020), 103056.
[58] Bran Knowles and John T. Richards. 2021. The sanction of authority: Promoting public trust in AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 262–271.
[59] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
[60] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82 (2017), 1–26.
[61] Justine Lacey, Mark Howden, Christopher Cvitanovic, and R. M. Colvin. 2018. Understanding and managing trust at the climate science–policy interface. *Nat. Clim. Change* 8, 1 (01 Jan. 2018), 22–28. DOI: https://doi.org/10.1038/s41558-017-0010-z
[62] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?" Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.
[63] Dong-Jin Lee, Moonkyu Lee, and Jaebeom Suh. 2007. Benevolence in the importer-exporter relationship: Moderating role of value similarity and cultural familiarity. *International Marketing Review* 24, 6 (2007), 657–677. https://doi.org/10.1108/02651330710832649
[64] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Hum. Fact.* 46, 1 (2004), 50–80.
[65] Scott LeeTiernan, Edward Cutrell, Mary Czerwinski, and Hunter G. Hoffman. 2001. Effective notification systems depend on user trust. In *Proceedings of the INTERACT Conference*. 684–685.
[66] Jie Leng and Jixia Wu. 2019. Integrity perceptions and behavior triggered by the hand-over-chest gesture: A semiotic perspective. *Language* 3 (2019).
[67] Roy J. Lewicki and Chad Brinsfield. 2015. Trust research: Measuring trust beliefs and behaviours. In *Handbook of Research Methods on Trust*. Edward Elgar Publishing.
[68] Q. Vera Liao and S. Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT'22)*. Association for Computing Machinery, New York, NY, 1257–1268. DOI: https://doi.org/10.1145/3531146.3533182
[69] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proc. ACM Hum.-comput. Interact.* 5, CSCW2 (2021), 1–45.
[70] Erina L. MacGeorge and Lyn M. Van Swol. 2018. *The Oxford Handbook of Advice*. Oxford University Press.
[71] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.
[72] Ronald Scott Marshall. 2003. Building trust early: The influence of first and second order expectations on trust in international channels of distribution. *Int. Bus. Rev.* 12, 4 (2003), 421–443.
[73] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 3 (1995), 709–734.
[74] Lynne McFall. 1987. Integrity. *Ethics* 98, 1 (1987), 5–20.
[75] John M. McGuirl and Nadine B. Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Hum. Fact.* 48, 4 (2006), 656–665.
[76] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. The impact of initial consumer trust on intentions to transact with a web site: A trust building model. *J. Strateg. Inf. Syst.* 11, 3-4 (2002), 297–323.
[77] David L. McLain and Katarina Hackman. 1999. Trust, risk, and decision-making in organizational change. *Public Administration Quarterly* 23, 2 (1999), 152–76. http://www.jstor.org/stable/40861778
[78] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 6 (2021), 1–35.
[79] Siddharth Mehrotra. 2021. Modelling trust in human-AI interaction. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 1826–1828.
[80] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More similar values, more trust? The effect of value similarity on trust in human-agent interaction. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, 777–783.
[81] Stephanie M. Merritt. 2011. Affective processes in human–automation interactions. *Hum. Fact.* 53, 4 (2011), 356–370.
[82] Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research? In *Proceedings of the Conference on Trust and Reliance in AI-Human Teams (TRAIT'22)*. 11.
[83] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLoS ONE* 15, 2 (2020).

[84] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G. Jentsch. 2013. Building appropriate trust in human-robot teams. In *Proceedings of the AAAI Spring Symposium Series*.
[85] Andrés Páez. 2019. The pragmatic turn in explainable artificial intelligence (XAI). *Minds Mach.* 29, 3 (2019), 441–459.
[86] Michael E. Palanski and Francis J. Yammarino. 2007. Integrity and leadership: Clearing the conceptual confusion. *Eur. Manag. J.* 25, 3 (2007), 171–184.
[87] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Hum. Fact.* 39, 2 (1997), 230–253.
[88] Christopher J. Peters. 1995. Foolish consistency: On equality, integrity, and justice in stare decisis. *Yale LJ* 105 (1995), 2031.
[89] David V. Pynadath and Stacy C. Marsella. 2005. PsychSim: Modeling theory of mind with decision-theoretic agents. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Vol. 5. 1181–1186.
[90] Denise Christine Rieser and Orlando Bernhard. 2016. Measuring trust: The simpler the better? In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2940–2946.
[91] Maria Riveiro and Serge Thill. 2021. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artif. Intell.* 298 (2021), 103507.
[92] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, 101–108.
[93] Matthew N. O. Sadiku, Sarhan M. Musa, and A. Ajayi-Majebi. 2021. *A Primer on Multiple Intelligences*. Springer.
[94] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Bus. Technol. J.* 31, 2 (2018), 47–53.
[95] Mary Steffel, Elanor F. Williams, and Jaclyn Perrmann-Graham. 2016. Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organiz. Behav. Hum. Decis. Process.* 135 (2016), 32–44.
[96] Micha Strack and Carsten Gennerich. 2011. Personal and situational values predict ethical reasoning. *Eur. J. Psychol.* 7, 3 (2011), 419–442.
[97] Gabriele Taylor and Raimond Gaita. 1981. Integrity. *Proc. Aristot. Soc., Supplem. Vol.* 55 (1981), 143–176.
[98] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[99] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 77–87.
[100] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. 2020. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*. 3–12.
[101] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020), 100049.
[102] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT'20)*. 272–283.
[103] Abdullah Aman Tutul, Ehsanul Haque Nirjhar, and Theodora Chaspari. 2021. Investigating trust in human-machine learning collaboration: A pilot study on estimating public anxiety from speech. In *Proceedings of the International Conference on Multimodal Interaction*. 288–296.
[104] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. 2009. Computing confidence values: Does trust dynamics matter? In *Proceedings of the Portuguese Conference on Artificial Intelligence*. Springer, 520–531.
[105] Mascha Van't Wout and Alan G. Sanfey. 2008. Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition* 108, 3 (2008), 796–803.
[106] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* 76 (2021), 89–106.
[107] Alan R. Wagner, Jason Borenstein, and Ayanna Howard. 2018. Overtrust in the robotic age. *Commun. ACM* 61, 9 (2018), 22–24.
[108] Alan R. Wagner, Paul Robinette, and Ayanna Howard. 2018. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Trans. Interact. Intell. Syst.* 8, 4 (2018), 1–24.
[109] Connie R. Wanberg and Paul M. Muchinsky. 1992. A typology of career decision status: Validity extension of the vocational decision status model. *J. Counsel. Psychol.* 39, 1 (1992), 71.

[110] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, 109–116.

[111] Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 318–328.

[112] Xinru Wang and Ming Yin. 2022. Effects of explanations in AI-assisted decision making: Principles and comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4 (2022). DOI: https://doi.org/10.1145/3519266

[113] T. B. Warrington, N. J. Abgrab, and H. M. Caldwell. 2000. Building trust to develop competitive Advantage in e-business Realtionship. *Competitiveness Review*, 10, 2 (2000), 160–168. https://doi.org/10.1108/eb046409

[114] Lawrence R. Wheeless and Janis Grotz. 1977. The measurement of trust and its relationship to self-disclosure. *Hum. Commun. Res.* 3, 3 (1977), 250–257.

[115] William Wilson. 2004. Suggestions to foster effective consultation within conservation. *Environments* 32, 2 (2004), 71.

[116] Michael Winikoff. 2017. Towards trusting autonomous systems. In *Proceedings of the International Workshop on Engineering Multi-agent Systems*. Springer, 3–20.

[117] Jingjun David Xu, Ronald T. Cenfetelli, and Karl Aquino. 2016. Do different kinds of trust matter? An examination of the three trusting beliefs on satisfaction and purchase behavior in the buyer-seller context. *J. Strat. Inf. Syst.* 25, 1 (2016), 15–31.

[118] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[119] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. *ACM Trans. Internet Technol.* 17, 1 (2017), 1–20.

[120] Qiaoning Zhang, Matthew L. Lee, and Scott Carter. 2022. You complete me: Human-AI teams and complementary expertise. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–28.

[121] 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 Conference on Fairness, Accountability, and Transparency*. 295–305.

[122] Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. 2021. Forward reasoning decision support: Toward a more complete view of the human-AI interaction design space. In *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter (CHItaly'21)*. 1–5.

[123] Yuhui Zhong, Bharat Bhargava, Yi Lu, and Pelin Angin. 2014. A computational dynamic trust model for user authorization. *IEEE Trans. Depend. Sec. Comput.* 12, 1 (2014), 1–15.

Received 21 July 2022; revised 4 July 2023; accepted 10 July 2023