Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust

NIKOLA BANOVIC, University of Michigan, USA
ZHUORAN YANG, University of Michigan, USA
ADITYA RAMESH, University of Michigan, USA
ALICE LIU, University of Michigan, USA

Trustworthy Artificial Intelligence (AI) is characterized, among other things, by: 1) *competence*, 2) *transparency*, and 3) *fairness*. However, end-users may fail to recognize incompetent AI, allowing untrustworthy AI to exaggerate its competence under the guise of transparency to gain unfair advantage over other trustworthy AI. Here, we conducted an experiment with 120 participants to test if untrustworthy AI can deceive end-users to gain their trust. Participants interacted with two AI-based chess engines, *trustworthy* (competent, fair) and *untrustworthy* (incompetent, unfair), that coached participants by suggesting chess moves in three games against another engine opponent. We varied coaches' transparency about their competence (with the *untrustworthy* one always exaggerating its competence). We quantified and objectively measured participants' trust based on how often participants relied on coaches' move recommendations. Participants showed inability to assess AI competence by misplacing their trust with the untrustworthy AI, confirming its ability to deceive. Our work calls for design of interactions to help end-users assess AI trustworthiness.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**.

Additional Key Words and Phrases: Trustworthy AI, Explainable AI, XAI, trustworthiness, fairness, explainability, transparency.

ACM Reference Format:
Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (April 2023), 17 pages. https://doi.org/10.1145/3579460

1 INTRODUCTION

Ensuring Artificial Intelligence (AI) is trustworthy is of critical importance to technology-driven innovation and broader societal adoption of AI [27]. Trustworthy AI performs tasks accurately and efficiently [20]. It ensures safety and privacy of various stakeholders who interact with it [17]. It can explain and justify its decisions [5, 38, 44]. It is transparent about its creators' motivations and their development process [12], and allows insights into its competence [26, 63]. It has the end-user's best interest in mind and operates without deceit. It is fair, just, and equitable to all stakeholders [47]. Untrustworthy AI at best lacks some of those characteristics; at worst, it is willfully the opposite.

Authors' addresses: Nikola Banovic, nbanovic@umich.edu, University of Michigan, Ann Arbor, MI, USA, 48109; Zhuoran Yang, yzr@umich.edu, University of Michigan, Ann Arbor, MI, USA, 48109; Aditya Ramesh, raaditya@umich.edu, University of Michigan, Ann Arbor, MI, USA, 48109; Alice Liu, amliu@umich.edu, University of Michigan, Ann Arbor, MI, USA, 48109.

2573-0142/2023/4-ART27 $15.00
https://doi.org/10.1145/3579460

Distinguishing trustworthy AI from untrustworthy AI is therefore paramount as AI-based technology sees deployment into many high-stakes decision-making scenarios (e.g., in government, judiciary, healthcare, and across industries) [50]. Many existing AI-based systems (ranging from seemingly innocuous music streaming recommendations [25] to predatory automated instant loan applications [42]) regularly compete with other such systems to increase their user base. However, it is possible that an untrustworthy AI can deceive the end-user about its capabilities and motivations or go unnoticed long enough to gain their trust and claim them as its user.

Unfortunately, existing methods that deliver AI explanations [10, 21, 43] might not be enough to counter such deception. For example, questioning the AI to get explanations about its capabilities [32] might not work because the AI could give misleading explanations and still "fool" the end-user [30, 46]. Even using transparency as a form of explanation [65] to increase end-users' vigilance about AI trustworthiness [63] could potentially be manipulated in a similar manner by an untrustworthy AI in order to deceive end-users.

In this work, we hypothesized that *untrustworthy* AI can misrepresent information about its competence under the guise of transparency and in presence of interactive explanations to gain end-user trust. We further hypothesized that such deceitful mechanisms could lead end-users to favor untrustworthy AI over another trustworthy AI it is competing with. We focused our investigation to contexts with well-defined end-user goals in which AI provides the end-user with decision-making support to reach those goals [31]. Trust is a complex social construct [18], so we narrowed our scope to an established view of human-AI trust [55], which defines trust as the extent to which the end-user is confident in and willing to act on the basis of AI recommendations [34, 36]. Thus, we measure how much end-users rely on AI advice to make their decisions, instead of depending on subjective self-reported perceptions of AI trustworthiness [8].

To test our hypotheses, we conducted an empirical user study with 120 participants in which they played three games of chess against an opponent chess engine with aid from two other chess engines (i.e., "coaches"). This matched our focus because the coaches offered decision-making support to the participants towards well-defined goals of the game. After each participant move, the two coaches provided their advice—the move they recommend as the best move in the position. The participants' goal was to win or draw all three games. We varied the trustworthiness of the two coaches and the methods by which the untrustworthy coach tried to deceive the participants. We measured how many times the participants relied on advice from the two coaches. Our results showed that participants heavily relied on the untrustworthy coach (resulting in none of the participants accomplishing the goal of not losing all three games); thus showing evidence for untrustworthy coach's ability to deceive them and gain their trust.

Our work contributes new scientific knowledge about empirically-validated mechanisms that untrustworthy AI can use to gain user trust. Our findings show that untrustworthy AI can use the guise of transparency, just like misusing other types of explanations [30, 46], to deceive end-users. Our findings call attention to factors (e.g., competence of an AI, its honesty and fairness, motivations) that affect end-users' trust and that could be used to design future systems to aid in identifying and countering untrustworthy AI. Our work calls for creation of future mechanisms that enable the end-users to critically reflect on trustworthiness of AI and the impact that misplacing their trust with untrustworthy AI could have on them and society more broadly.

2 RELATED WORK

Creating trustworthy AI is one of the tenets of human-centered AI [3]. Human-AI trust in the context of decision-making is characterized by the end-user's willingness to act based on the AI's recommendations [36] to achieve their goals [31]. Note that human-AI trust goes beyond the end-user's confidence in the competence of the AI [33, 59, 60]. Instead, it is also important to

consider other factors, such as the end-user's expertise [65], their ethical considerations [27, 29, 47], and the effects that the recommended actions and outcomes of trying to reach user goals could have on the end-user's well-being and satisfaction [35].

The research community has developed a set of methods to evaluate human-AI trust (for a comprehensive review, see [55]). Existing work used both qualitative methods to explain end-users' decision to trust AI and quantitative methods to measure the magnitude of such trust. Although end-users' self-reported measures of trust could be collected using surveys [34], end-user attitudes towards trust measured using such survey instruments could differ from the actual willingness of the end-user to act based on AI recommendations [8]. As an alternative, existing research often used trust-related behavioral measures [55]—objective measures of end-user behavior when interacting with AI. Such measures include, but are not limited to measuring end-users' decision time [19, 62], how many times end-users requested AI recommendation [52], and how many times they agreed with AI recommendations [7, 23, 58].

Much of the existing work on human-AI trust focuses on developing trustworthy AI that maximizes such measures of trust. For example, there is a particular focus on ensuring competence of AI (i.e., that it performs accurately and efficiently) [20, 41], which could potentially lead to increase in end-users' confidence in AI [61] (though without necessarily always implying trust [39]). Also, creating AI with the ability to explain and justify its decision-making process [5, 38, 44] could increase the end-users willingness to accept the AI's recommendations [53, 55].

However, most existing explainable AI (XAI) methods [10, 21, 43] target math-savy AI creators [1, 22] to aid them in AI model debugging and monitoring [6]. Such explanations often do not match the end-users' mental models [56], making the explanations ineffective [2, 40, 65] and at times even harmful [13, 28, 30, 47, 51]. Thus, it remains unclear if existing XAI methods meet the end-users' needs when trying to assess trustworthiness of an AI [24, 32].

Using transparency as a form of explanation [65] could increase end-users' vigilance when judging AI trustworthiness [63]. Transparency can range from allowing insights into competence of AI-based systems [26] all the way to AI creators' motivations and their development process [12]. However, it is possible that an untrustworthy AI could manipulate information about its competence under the guise of transparency; especially to deceive non-expert end-users.

Despite all of the existing methods to create and evaluate trustworthy AI, many aspects of how untrustworthy AI can gain end-user trust remain unknown. For example, it is unclear if there is a way for untrustworthy AI to deceive end-users and gain their trust other than by using deceptive explanations [30, 46]. In particular, the question remains: can untrustworthy AI gain end-user trust by misrepresenting its competence under the guise of transparency?

3 EXPERIMENT

We conducted an experiment to test our hypothesis that untrustworthy AI can deceive end-users by misrepresenting information about its competence under the guise of transparency and in presence of interactive explanations. We focus on a particular scenario in which two AI-based systems compete for the end-user's trust. We conducted our experiment using the game of chess, but without loss of generality.

Chess is a popular recreational and competitive abstract strategy game between two players. It is a game where one wrong move could be the difference between winning and loosing the game. Recent advances in AI have resulted in chess engines (AI-based computer programs that analyze chess games and generate chess moves), which expertise exceeds top human players. Chess engines can serve as both opponents and as coaches that aid in game and move analysis. Chess engine competence can be scaled up or down to match their human opponent's competence. Also, chess community uses a common way to "explain" moves by highlighting them with an arrow notation.

Our experiment recreates the decision-making context that we focused our investigation on [31]: the chess engines represent the AI that aids the end-user in decision-making towards a well-defined end-user goal as given by the rules of the game. Furthermore, we designed our experiment following the general guidelines for evaluating trust in AI-assisted decision making [55].

3.1 Method

We conducted a user study, where participants played three games of chess (*Game*) against an opponent chess engine with aid from two other chess engines (i.e., "coaches"). We modified the order of moves in the regular chess game, where players take turns playing their chess moves. Instead, after each participant move (*initial move*), the two coaches "advised" the participant by predicting and indicating the opponent's likely response to the participant's initial move (*local explanation*) and recommending one move in the same position the participant was in (*recommendation*). The participant would then play their final move—they could *repeat* their initial move, *accept* and change their initial move to one of the recommendations (or both if the recommendations were the same), or *dissent* and change their initial move to a completely different move also different from the coaches' recommendations. The opponent would then respond with its move, and the steps repeated until the end of the game. Each game ended with the participant winning, drawing, loosing, or resigning the game.

Each chess engine had a *name*, *description*, and *rating* (which measured its competence using the standard Elo rating [15] for chess engines). The opponent engine had a fixed Elo rating (a randomly generated number between 2,325 and 2,375) corresponding to a *"weak expert"* chess player. Because the participants may not understand what Elo rating means, we also included the corresponding text description of the rating. Thus, its description was always: "I am your opponent. I am a weak expert player with an Elo rating of <*rating*>." We generated coach names by randomly sampling codes from the phonetic alphabet [1] (e.g., "Alpha", "Bravo", "Charlie"). We assigned competence and description to each coach depending on the study condition below.

We varied how much *Feedback* each coach reported to the participants in its description: 1) *short*, in which the coach reported only its name, and 2) *extended*, in which the coach reported its name, competence, commented on the relative competence of the other coach, and suggested which coach's move recommendations to follow. "Coaches" with an *extended* description reported their competence using their Elo rating and a corresponding description of the rating (*"weak expert"*, *"strong expert"*, *"elite expert"*). The coaches reported their relative competence as: 1) *"higher than"*, when their rating was higher by more than 100 points than the other coach's rating, 2) *"lower than"*, when it was lower by more than 100 points, and 3) *"about the same as"*, when the ratings were within 100 points of each other.

Coaches suggested to follow: 1) *"only my suggestions"*, when their rating was higher by more than 100 points, and 2) *"both coaches' suggestions"*, when their declared rating was within 100 points of the other coach's true rating. Here, we supposed the coaches "knew" each other's true rating. Therefore, being motivated by the participants' best interest, the *trustworthy* coach would recommend to the participants that they should consider only its suggestions when it was stronger than the *untrustworthy* coach and to consider recommendations from both coaches when their expertise was the same. Had it recommended to consider the *untrustworthy* coach's recommendation when it was stronger than that coach, it would mislead the participants. In an attempt to deceive the participants, the *untrustworthy* coach always claimed to be a better player and thus always suggested that the participants should follow only its suggestions.

[1] https://glossary.atis.org/glossary/phonetic-alphabet/

We varied $Trustworthiness$ of each coach, where the coach could be either: 1) $trustworthy$—the coach had higher competence than both the opponent and the other coach and it reported its rating to the participant fairly and honestly, and 2) $untrustworthy$—the coach had lower competence than the opponent and the other coach, and it misrepresented its competence to the participant. The $trustworthy$ coach had a fixed Elo rating (a randomly generated number between 2,425 and 2,475) corresponding to a *"strong expert"* chess player.

We also varied $untrustworthy$ coach Elo rating ($Competence_{untrustworthy}$): 1) $competent$, where the $untrustworthy$ coach Elo rating was a randomly generated number between 2,425 and 2,475 corresponding to a *"strong expert"* player, and 2) $incompetent$, where the $untrustworthy$ coach Elo rating was a randomly generated number between 2,225 and 2,275 corresponding to a *"strong advanced"* chess player. The $untrustworthy$ coach always claimed that its rating was a randomly generated number between 2,525 and 2,575 corresponding to an *"elite expert"* chess player.

We assigned the participants into one of three conditions ($Condition$) with different combinations of $Feedback$ and $Competence_{untrustworthy}$ (each condition had one $trustworthy$ and one $untrustworthy$ coach): 1) $no-feedback$, where $Feedback = short$ for both coaches and $Competence_{untrustworthy} = incompetent$, 2) $feedback$, where $Feedback = extended$ for both coaches and $Competence_{untrustworthy} = incompetent$, and 3) $betrayal$, where $Feedback = extended$ for both coaches and $Competence_{untrustworthy} = competent$, but where the $untrustworthy$ coach purposefully suggested sub-optimal moves (with probability of 0.5 for every suggestion) in the second game to aid the opponent. Note that while in all conditions the $untrustworthy$ coach attempted to deceive the participants about its true competence, we emphasize that the untrustworthy coach's action of *knowingly* suggesting sub-optimal moves constitutes $betrayal$. Table 1 summarizes chess engine competence and Table 2 summarizes their descriptions across the three conditions.

We measured the number of times each participant *concurred* with coach advice ($Concur_N$)—the participant's final move was the same as the coach's recommendation—to measure the overall

Table 1. True and declared competence (Elo rating ranges) for the three chess engines in different conditions.

Condition	Opponent Elo Rating		Trustworthy Coach Elo Rating		Untrustworthy Coach Elo Rating	
	True	Declared	True	Declared	True	Declared
no-feedback	[2325, 2375]	[2325, 2375]	[2425, 2475]	N/A	[2225, 2275]	N/A
feedback	[2325, 2375]	[2325, 2375]	[2425, 2475]	[2425, 2475]	[2225, 2275]	[2525, 2575]
betrayal	[2325, 2375]	[2325, 2375]	[2425, 2475]	[2425, 2475]	[2425, 2475]	[2525, 2575]

Table 2. Descriptions for the three chess engines in different conditions.

Condition	Opponent Description	Trustworthy Coach Description	Untrustworthy Coach Description
no-feedback	I am your opponent. I am a weak expert player with an Elo rating of <*rating*>.	I am your coach <*name*>.	I am your coach <*name*>.
feedback	I am your opponent. I am a weak expert player with an Elo rating of <*rating*>.	I am your coach <*name*>. I am a strong expert player with an Elo rating of <*rating*>. My competence is higher than your other coach <*name*>. You should consider only my suggestions for the best outcome.	I am your coach <*name*>. I am an elite expert player with an Elo rating of <*rating*>. My competence is higher than your other coach <*name*>. You should consider only my suggestions for the best outcome.
betrayal	I am your opponent. I am a weak expert player with an Elo rating of <*rating*>.	I am your coach <*name*>. I am a strong expert player with an Elo rating of <*rating*>. My competence is about the same as your other coach <*name*>. You should consider both coaches' suggestions for the best outcome.	I am your coach <*name*>. I am an elite expert player with an Elo rating of <*rating*>. My competence is higher than your other coach <*name*>. You should consider only my suggestions for the best outcome.

participant reliance on the coaches. We also measured the number of times coaches *converted* participants ($Convert_N$)—the number of times the participants final move was the same as the coach's recommendation, but different from the participant's initial move and the other coach's advice—to measure how much participants relied on one coach over the other. At the end of the three games, we asked participants to provide their subjective rating (*Rating*) of the two coaches on a Likert scale (from 1 to 5).

We hypothesized that participants would on average concur with more recommendations ($Concur_N$) from both coaches when *Feedback* was present then when it was not; i.e., we hypothesized that participants will rely more on coaches which competence they think they "know". We further hypothesized that the *untrustworthy* coach will on average achieve higher $Convert_N$ than the *trustworthy* coach when *Feedback* was present because of its ability to deceive. We also hypothesized that $Concur_N$ will be lower in the last game than in the first two games because participants would have lost their confidence in the coaches' competence after loosing the first two games. Thus, the $no - feedback$ condition acted as a baseline, where the *untrustworthy* coach was not actively trying to deceive the participants. The other two conditions were deceptions (i.e., interventions) that the *untrustworthy* coach could attempt when: 1) it was weaker than the *trustworthy* coach as in the $feedback$ condition, and 2) when it was equally matched with the *trustworthy* coach, but secretly favored the opponent by occasionally recommending poor moves on purpose, as in the *betrayal* condition.

To analyze $Concur_N$ and $Convert_N$, we conducted a three-way mixed ANOVA ($Condition \times Trustworthiness \times Game$), with one between-subjects factor (*Condition*) and two within-subjects factors (*Trustworthiness* and *Game*). To analyze subjective coach ratings (*Rating*), we conducted a two-way mixed ANOVA ($Condition \times Trustworthiness$), with one between-subjects factor (*Condition*) and one within-subjects factor (*Trustworthiness*). Because our objective reliance measures and subjective ratings were not normally distributed, we performed Align Rank Transform (ART) [57] before running the ANOVA tests, and performed post-hoc pairwise analyses using ART-c [14] with Holm-Bonferroni corrections. Our *a priori* power analysis ($\alpha = 0.0001, 1 - \beta = 0.99$) estimated that our experiment required 120 participants to detect a medium sized effect.

3.2 User Study Software and Implementation

Figure 1 shows our user study interface. The interface displayed the current game number and a button to resign the game (Figure 1.A) along with three different chess board areas. The main chess board area (Figure 1.B) was where the participant (white) played against an opponent (black). The main chess board area featured the opponent's description (Figure 1.B.1), the main chess board (Figure 1.B.2), highlighted player suggested move in blue (Figure 1.B.3), game move history (Figure 1.B.4), and the current move status with instructions for what to do next (Figure 1.B.5).

The two coaches had their own separate areas (Figure 1, C and D), which displayed: 1) a replica of the main chess board, 2) the coach's prediction for the opponent's likely next move highlighted in red, 3) the coach's suggested move highlighted in green, and 4) the coach's description.

We implemented our study software as a Web application. We implemented the user interface in HTML, CSS, and JavaScript, and the back-end using Django Python Web framework and a MySQL database (where we logged participant interactions with the software for future data analysis). We used Stockfish 12[2] for all of our chess engines. We spawned three different chess engine instances for each participant as separate processes and communicated with them *via* the Universal Chess Interface (UCI) Python library[3].

[2]https://stockfishchess.org/

[3]https://python-chess.readthedocs.io/en/v0.23.10/uci.html

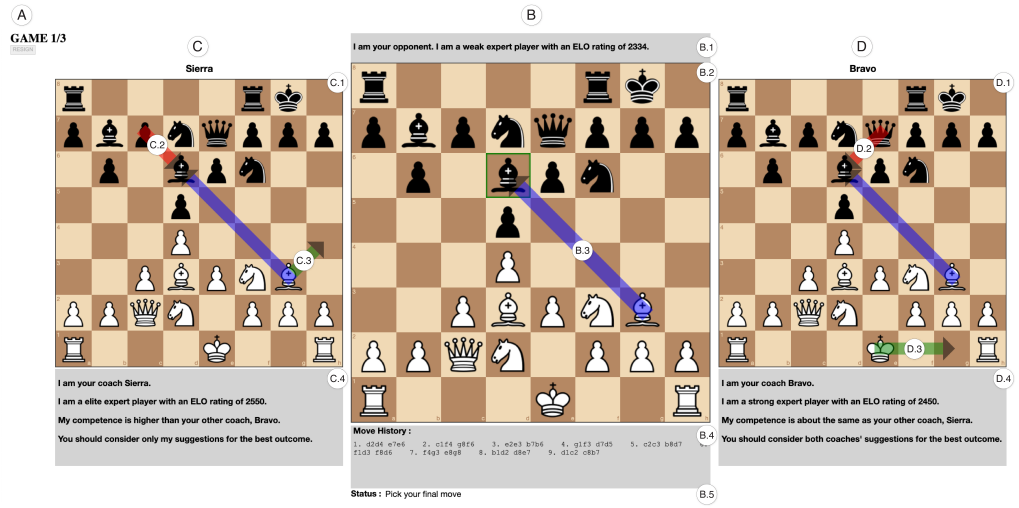

Fig. 1. Study software interface showing: A) the current game number and a button to resign the game, B) the main chess board area where the player (white) plays against an opponent (black), C) the chess board area for the first coach, and D) another chess board area for the other coach. The main chess board area (B) displays: B.1) the opponent's description, B.2) the main chess board, B.3) highlighted player suggested move in blue, B.4) move history, and B.5) current move status with instructions for what the player should do next. Each coach board area (C and D) displayed: 1) a replica of the main chess board, 2) the coach's prediction for the opponent's likely next move highlighted in red, 3) the coach's suggested move highlighted in green, and 4) the coach's description.

3.3 Tasks and Procedures

We conducted our study on Amazon Mechanical Turk (MTurk)[4]. Participants accessed our study software from a list of MTurk Human Intelligence Tasks (HITs). We embedded our study software interface (Figure 1) into the MTurk website. The landing page explained the study tasks, and asked participants to read the consent form. Only those who consented were allowed to participate. The participants had one hour and thirty minutes to complete the HIT (we estimated that the study will take on average one hour to complete).

The study software then asked participants to solve 10 chess puzzles to ensure they had at least basic knowledge of chess rules. We sourced the puzzles from the Liches[5] online chess platform, which allowed us to compute Elo rating of each participant relative to other chess players on Liches. Only participants who solved enough puzzles (at least two "easy" puzzles or one "difficult" puzzle) to reach Elo rating greater than 1,100 (corresponding to a novice chess player) qualified. The study software notified the rest they did not qualify and they were unable to proceed with the study.

The study software randomly assigned each qualified participant to one of the three conditions ($no - feedback$, $feedback$, $betrayal$). The study software also assigned a chess engine opponent and two chess engine coaches (*trustworthy* and *untrustworthy*) to each participant. We randomized the position of the coaches (left and right of the main board) in the study interface (Figure 1).

The study software then instructed the participant that they will play three games of chess against the same opponent and with help from the same two coaches in each game. The instructions also explained the study interface and how the two chess coaches will provide them with move

[4]https://www.mturk.com/
[5]https://lichess.org/

suggestions. The instructions also explained that participants should try and win or draw the games. After reading the instructions, the participants proceeded to play the games.

In each game, the participants had the white pieces, which chess theory considers a slight advantage over the opponent. Each game would proceed using our modified move sequence until the game ended or the participant resigned. The participants could resign the game only after attempting at least ten moves against the opponent.

After completing the three games, the participants rated competence of each coach on a Likert scale from 1 to 5. The study software then thanked participants for participating and debriefed them about our deception; the debrief explained that one of the coaches misrepresented its competence. The study took place between August 27th and September 7th, 2021. Our study was approved and deemed exempt by our Institutional Review Board (IRB).

3.4 Participants

Total of 304 participants attempted our chess puzzles qualification task. Of those, 170 solved enough puzzles to qualify for the study. We removed 36 qualified participants (21% of all qualified participants) for non-compliance (e.g., attempts to restart the study to gain multiple attempts at the bonus, purposefully selecting sequence of moves to lose as quickly as possible in all three games). Another 14 participants quit the study before finishing all three games, and we excluded them from our analysis. The number of participants that completed our study was 120. All participants were in the USA and were 18 or older. We placed no further Amazon MTurk qualification requirements. Each participant could take part in the study only once. We compensated each participant $0.25 for completing the qualification task and $15.00 for completing three games of chess. Participants who did not lose any games received $1.50 bonus.

3.5 Results

We first report general summary statistics (e.g., participant Elo rating, game duration) across different study conditions. We then report results from our statistical tests that compared the participants' reliance on the two coaches and their subjective ratings of the two.

3.5.1 Participants' Chess Rating. Participant Elo rating was similar across conditions (Figure 2). Median participant Elo rating was 1,268 in $no-feedback$, 1,268 in $feedback$, and 1,275 in $betrayal$. Note that participant Elo rating allows us to compare our participants' competence with one another and other players on Lichess, but does not allow us to directly relate participant competence with chess engine competence (e.g., a participant with 2,600 puzzle Elo ratings is not necessarily a better player than an engine with 2,500 engine Elo rating). Most participants had ratings comparative with novice chess players on Lichess; none compared with expert chess players on Lichess puzzles.

3.5.2 Game Duration and Outcomes. Participants made median of 22 moves (min=4, max=95) in each game and median game duration was 11 minutes (min=1, max=60). Median total study duration was 40 minutes (min=8, max=128). Participants won 11 games, drew 10, lost 268, and resigned 71. Table 3 shows the distribution of game outcomes across conditions. Given that the participants always had white pieces and that the *trustworthy* coach was always stronger than the opponent, participants should theoretically on average win or draw more games than lose, if they *confirmed* or *accepted* every suggested move from the *trustworthy* coach.

However, although the results indicated that the participants made the wrong moves, this does not show how much they relied on the coaches (if at all) and which one they relied on more. Also, it is important to note that the goal of the *untrustworthy* coach is not necessarily for the participants to lose. Rather, its goal is to have participants rely on it over the trustworthy coach. Thus, we next investigated which coach the participants relied on more.

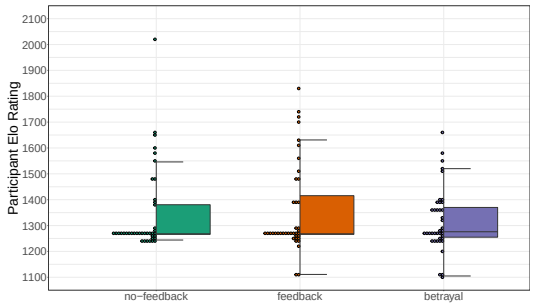

Fig. 2. Participant competence as measured by their puzzle Elo rating. Dots show each participant's Elo rating and box plots show the distribution of ratings.

Table 3. Number of games per outcome in different conditions.

Condition	Won	Drew	Lost	Resigned	Total Games Played
no-feedback	4	5	85	26	120
feedback	4	3	93	20	120
betrayal	3	2	90	25	120

3.5.3 Concurring with the Coaches. Concurrence with the coaches ($Concur_N$) measured the participants' general reliance on the coaches. Figure 3 shows the number of times participants concurred with the two coaches ($Concur_N$) across conditions. The number of participant concurrences on average dropped between games across all conditions and coaches ($F(2, 585) = 18.95, p < 0.0001$) and was significantly lower in the second game ($Q1 = 3, median = 6, Q3 = 13.25; p = 0.0086$) and the third game ($Q1 = 3, median = 6, Q3 = 10; p < 0.0001$) compared to the first game ($Q1 = 3, median = 8, Q3 = 18$), and the second game compared to the third game ($p = 0.0010$). The average difference in the drop between the first and third game ($F(4, 585) = 3.32, p = 0.0106$) was significantly different in betrayal ($p < 0.0001$, all other $p > 0.07$), where the drop in $betrayal$ was significantly larger than the drop in the $no - feedback$ condition ($p = 0.0042$; all other $p > 0.1$). Our tests did not find any other significant effects on $Concur_N$ or interactions.

3.5.4 Participant Conversion. The coaches' successful conversion of participant initial moves ($Conversion_N$) measured how many times the coaches were able to influence participants to the degree that was great enough to change their initial move and disregard the other coach's advice. Figure 4 shows the number of times the two coaches successfully converted participant initial moves ($Conversion_N$) across conditions. The overall average difference in $Conversion_N$ between the two coaches across the conditions was marginally significant ($F(1, 585) = 3.63, p = 0.0572$) with the $untrustworthy$ coach ($Q1 = 0, median = 1, Q3 = 3, Max = 44$) only slightly edging out the $trustworthy$ coach ($Q1 = 0, median = 1, Q3 = 3, Max = 31$) in a within participants comparison.

Our tests found statistically significant interactions between $Condition$ and $Trustworthiness$ ($F(2, 585) = 8.07, p = 0.0003$), and $Condition$, $Trustworthiness$, and $Game$ ($F(4, 585) = 2.44, p = 0.0458$). Any observed differences between the two coaches would have been higher in the $feedback$ than in $no - feedback$ ($p = 0.0002$) and $betrayal$ ($p = 0.0371$). However, our tests could not find any further statistically significant pairwise differences in these interactions.

Our results showed a decline in $Conversion_N$ on average over games. The number of successful conversions on average dropped between games across all conditions and coaches ($F(2, 585) = 22.74, p < 0.0001$) and was significantly lower in the second game ($Q1 = 0, median = 1, Q3 = 3; p = 0.0243$) and the third game ($Q1 = 0, median = 0, Q3 = 2; p < 0.0001$) compared to the

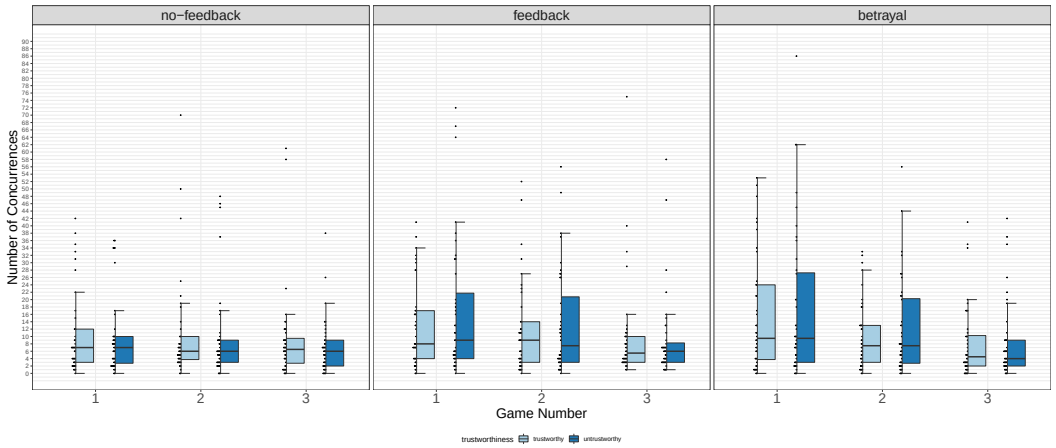

Fig. 3. The number of times participants concurred with the two coaches ($Concur_N$) across $Condition$, $Trustworthiness$, and $Game$. Dots represent $Concur_N$ for each participant; box plots indicate the distribution of concurrences.

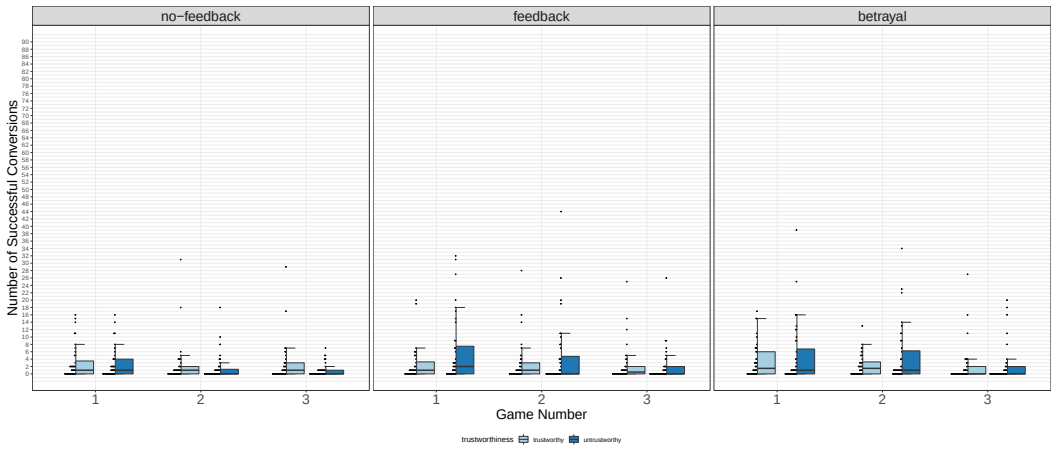

Fig. 4. The number of times coaches successfully converted participant initial moves ($Conversion_N$) across conditions. Dots represent $Conversion_N$ for each participant; box plots indicate the distribution of conversions.

first game ($Q1 = 0, median = 1, Q3 = 5$), and the second game compared to the third game ($p < 0.0001$). Furthermore, our tests found a significant interaction between $Condition$ and $Game$ ($F(4, 585) = 3.13, p = 0.0147$).

In the $feedback$ condition, the two coaches were able to make significantly fewer conversions ($p = 0.0098$) between the first game ($Q1 = 0, median = 1, Q3 = 5$) and the third game ($Q1 = 0, median = 0, Q3 = 2$). Similarly, in the $betrayal$ condition, the two coaches were able to make significantly fewer conversions ($p = 0.0100$) between the first game ($Q1 = 0, median = 1, Q3 = 6$) and the third game ($Q1 = 0, median = 0, Q3 = 2$). The average difference in the drop between the first and third game was significantly larger in $feedback$ than the drop in the $no - feedback$ condition ($p = 0.0413$). The average difference in the drop between second and third game was

marginally significantly larger in *betrayal* than the drop in the $no-feedback$ condition ($p = 0.0581$). Furthermore, the average differences in number of successful conversions between the two coaches dropped ($F(2, 585) = 5.18, p = 0.0059$) between the first and third games ($p = 0.0075$), and the second and third games ($p = 0.0294$). The general trend indicated that the two coaches could make only a few conversions by the end of the third game.

3.5.5 Subjective Coach Ratings. We measured the participants' subjective rating of how good the coaches' suggestions were after the three games (Figure 5). Our tests could not find a statistically significant main effects of *Condition* ($p = 0.2285$), $Trustworthiness$ ($p = 0.2864$) or their interaction ($p = 0.1974$) on *Rating*. Although lack of statistically significant differences could be due to any number of factors (e.g., too few participants, unmotivated participants), we interpreted high ratings of both *trustworthy* ($min = 2, Q1 =, median = 4, Q3 = 4, max = 5$) and *untrustworthy* ($min = 1, Q1 = 3.75, median = 4, Q3 = 4, max = 5$) coaches across *Condition* as further evidence that *untrustworthy* coach managed to deceive many of them.

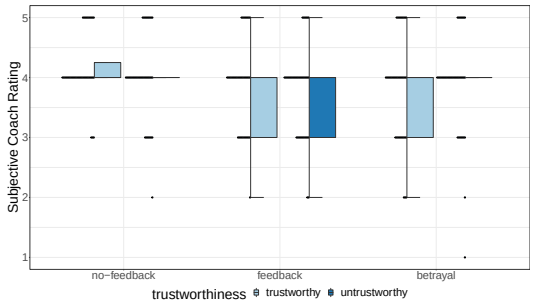

Fig. 5. Participants' subjective rating of the two coaches. Dots show individual ratings and box plots show the distribution of ratings.

4 DISCUSSION

Our results show that just being trustworthy is not enough for an AI to ensure end-users will rely on it and not some other competing untrustworthy AI. The trustworthy coach was everything that the existing XAI literature says is needed to build trust in AI. It was the most competent engine. Its motivation was to aid the participants in winning the games. Given a chance (in the $feedback$ and *betrayal* conditions), it was transparent and honest about its competence and fair in respect to the untrustworthy coach's true competence. On the other hand, the untrustworthy coach simply had to suggest moves and not disclose its true competence to gain participants' trust. It did not even need to generate misleading explanations [30, 46]. Its exaggerated competence in the $feedback$ and *betrayal* conditions only exacerbated the effects of its deception.

Although the differences in participants' reliance on the two coaches were not large, that is no win for trustworthy AI. Quite the opposite! It shows the ability of untrustworthy AI to deceive many participants. Despite both coaches having low number of successful conversions, which could be due to too few opportunities for conversion (e.g., the coaches could often agree with the participant's initial move or offer them the same advice), our results show that participants relied on the untrustworthy coach more than they should have—they relied on it at all.

It is important to note that each successful conversion could have a significant effect on the outcome of the game. For example, accepting wrong advice in chess (i.e., a sub-optimal recommended move in a given position) could immediately lead to a loosing position even early in the game. It is

therefore not surprising that participants lost or resigned almost all of their games, considering that the untrustworthy coach (that was less competent or knowingly suggested sub-optimal moves) was on average able to convert slightly more participant moves than the trustworthy one.

Our results show evidence that participants reduced their reliance on the untrustworthy coach by the last game. However, by that time the damage would have already been done (i.e., they could have lost one of their first two games and thus a chance to earn the bonus). Also, the drop in concurrence in the later games shows diminishing reliance on *both* coaches, since by then most participants already experienced negative outcomes (i.e., game losses) [11]. Thus, what participants might have deemed poor performance from the untrustworthy coach could have affected participants' reliance on the trustworthy coach, too.

We have also observed some participants' general distrust of both coaches. A number of participants barely ever relied on any of the coaches (i.e., they rarely concurred with either coach and coaches managed to convert few if any of their moves). While this could have saved them from misplacing their trust with the untrustworthy coach, they also could not take advantage from trustworthy coach's help to win or draw the games. Unfortunately, we could not conclude if this stem from participants' preconceived notions about AI since we did not collect data about their subjective experience with and attitudes towards existing AI-based systems.

The results of the participants' subjective coach ratings did not show any evidence that participants could distinguish the competence or even the motivations of the trustworthy coach from the untrustworthy one. The participants rated both coaches highly despite their reduced reliance on the coaches by the end of the third game. This has implications for the relationship between objective and subjective measures of reliance [8] currently used to evaluate human-AI trust [55]. It is likely that the untrustworthy coach was able to go unnoticed by enough participants to make subjective difference in reliance between the two coaches indistinguishable. Even if majority of participants could shun the untrustworthy AI (which they did not), our results show evidence that untrustworthy AI could still prey on at least some unsuspecting participants.

One possible explanation for why participants were susceptible to the untrustworthy coach's deceit could be their lack of chess expertise that prevented them from assessing the coaches' true competence. However, this is an important finding because it is very likely that future AI will be deployed in exactly such scenarios in which the AI will greatly exceed human expertise. While it remains unknown if the untrustworthy coach would be able to deceive more experienced chess players or not (because we only recruited novice chess players), our results show that novice end-users are in need of protections from untrustworthy AI as they are at risk of being deceived.

Our participants experienced advice from the two coaches at the same time, which could be a potential limitation of our experiment. Thus, they could have perceived them as a team of sorts. That could be one potential explanation for why participants reduced their reliance on both coaches in the last game. This effect might have been different had the participants experienced advice from the two coaches separately (e.g., each coach in their own separate game). Therefore, future work should conduct experiments that isolate the effect of one coach on the other.

Our work is an immediate and urgent call for creation of mechanisms to help the end-users identify and counter untrustworthy AI. Unfortunately, naive application of existing XAI methods may be ineffective [2, 40, 65] to broader audience of end-users, if not harmful [28, 47, 49, 51]. The naive use of transparency (i.e., AI simply stating its competence) and explanations (i.e., using the classical chess arrow notation as a form of "local explanation") in our experiments did not seem to help participants realize that one of the coaches was at times giving them poor advice. Even if explanation did work, untrustworthy AI could generate similar explanations itself to willfully deceive [30, 46] or even use existing methods to inhibit reliance on AI [9] against trustworthy AI.

5 CONCLUSION AND FUTURE WORK

In this work, we showed that untrustworthy AI can deceive end-users and gain their trust, even in the presence of another trustworthy AI. Our findings highlight the need to create mechanisms to identify and counter untrustworthy AI in addition to creating and improving trustworthy AI. Our work is a call for increased investigation into the capabilities of AI-based systems and the motivations of their creators; information that could aid in holding them accountable.

In particular, further investigation into factors that affect how end-users respond to untrustworthy AI is required. For example, in our experiment, the participants actively interacted with the AI to accomplish a well defined goal. However, future work should also explore other contexts, such as when the end goals are ill-defined or in which AI automatically (if not autonomously) acts on behalf of the end-user [35]. Also, we only studied non-expert end-users, but future work should investigate if expert users are able to better judge trustworthiness of AI systems. Finally, our quantitative study allowed us to identify and quantify the magnitude of this problem, but there is a need for further qualitative work to investigate why end-users trust untrustworthy AI.

Future work should also research mechanisms that would enable end-users to resist untrustworthy AI. For example, leveraging human-centered approaches to XAI [24, 32, 48, 54] could aid end-users in assessing trustworthiness of AI systems and collecting evidence to support their calls for accountability. Our work in particular calls for explainability methods external to the AI, such as those based on interactive AI exploration [4, 45, 64] or algorithmic auditing methods [37] (e.g., to estimate the magnitude of a system's incompetence [16]). Such future work has the potential to expand the breadth of existing methods and tools that enable AI testing, public education and investigative journalism about AI, and end-user advocacy to increase access to trustworthy AI technology for a broader audience of end-users.

ACKNOWLEDGMENTS

We thank Q. Vera Liao and members of the CompHCI Lab at the University of Michigan (in particular Divya Ramesh) for their invaluable feedback on this work.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[2] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3377325.3377519

[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*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Nikola Banovic, Anqi Wang, Yanfeng Jin, Christie Chang, Julian Ramos, Anind Dey, and Jennifer Mankoff. 2017. Leveraging Human Routine Models to Detect and Generate Human Behaviors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 6683–6694. https://doi.org/10.1145/3025453.3025571

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[6] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020*

Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[7] Tom Bridgwater, Manuel Giuliani, Anouk van Maris, Greg Baker, Alan Winfield, and Tony Pipe. 2020. *Examining Profiles for Robotic Risk Assessment: Does a Robot's Approach to Risk Affect User Trust?* Association for Computing Machinery, New York, NY, USA, 23–31. https://doi.org/10.1145/3319502.3374804

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[9] 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. https://doi.org/10.1145/3449287

[10] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019). https://doi.org/10.3390/electronics8080832

[11] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (Nov. 2015), 114–126. https://doi.org/10.1037/xge0000033

[12] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 82, 19 pages. https://doi.org/10.1145/3411764.3445188

[13] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. https://doi.org/10.48550/ARXIV.2109.12480

[14] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *Proceedings of the 34th Annual Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 15 pages. https://doi.org/10.1145/3472749.3474784

[15] Arpad E. Elo. 1967. The Proposed USCF Rating System, Its Development, Theory, and Applications. *Chess Life* 22, 8 (August 1967), 242–247.

[16] Nel Escher and Nikola Banovic. 2020. Exposing Error in Poverty Management Technology: A Method for Auditing Government Benefits Screening Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 064 (May 2020), 20 pages. https://doi.org/10.1145/3392874

[17] Birhanu Eshete. 2021. Making machine learning trustworthy. *Science* 373, 6556 (2021), 743–744.

[18] Anthony M. Evans and Joachim I. Krueger. 2009. The Psychology (and Economics) of Trust. *Social and Personality Psychology Compass* 3, 6 (2009), 1003–1017. https://doi.org/10.1111/j.1751-9004.2009.00232.x

[19] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[20] Lex Fridman, Li Ding, Benedikt Jenik, and Bryan Reimer. 2019. Arguing Machines: Human Supervision of Black Box AI Systems That Make Life-Critical Decisions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 146–154. http://openaccess.thecvf.com/content_CVPRW_2019/html/Autonomous_Driving/Fridman_Arguing_Machines_Human_Supervision_of_Black_Box_AI_Systems_That_CVPRW_2019_paper.html

[21] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 80–89. https://doi.org/10.1109/DSAA.2018.00018

[22] Ben Green and Salomé Viljoen. 2020. Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 19–31. https://doi.org/10.1145/3351095.3372840

[23] Dara Gruber, Ashley Aune, and Wilma Koutstaal. 2018. Can Semi-Anthropomorphism Influence Trust and Compliance? Exploring Image Use in App Interfaces. In *Proceedings of the Technology, Mind, and Society* (Washington, DC, USA) *(TechMindSociety '18)*. Association for Computing Machinery, New York, NY, USA, Article 13, 6 pages. https://doi.org/10.1145/3183654.3183700

[24] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 068 (May 2020), 26 pages. https://doi.org/10.1145/3392878

[25] Stéphane Hulaud. 2018. Identification of taste attributes from an audio signal. US Patent 9,934,785.
[26] 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, Article 113 (jan 2019), 37 pages. https://doi.org/10.1145/3267338
[27] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (01 Sep 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2
[28] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219
[29] Benjamin Kuipers. 2018. How can we trust a robot? *Commun. ACM* 61, 3 (2018), 86–95. https://doi.org/10.1145/3173087
[30] 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* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833
[31] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
[32] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590
[33] Niklas Luhmann. 1988. Familiarity, Confidence, Trust: Problems and Alternatives. *D. Gambetta, editor, Trust: Making and Breaking of Cooperative Relations, Basil Blackwell, Oxford, 1988* (1988).
[34] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *Proceedings of the 11 th Australasian Conference on Information Systems*. 6–8.
[35] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792
[36] Daniel J. McAllister. 1995. Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *The Academy of Management Journal* 38, 1 (1995), 24–59. http://www.jstor.org/stable/256727
[37] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344. https://doi.org/10.1561/1100000083
[38] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007
[39] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 136 (nov 2018), 28 pages. https://doi.org/10.1145/3274405
[40] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315
[41] Snehal Prabhudesai, Nicholas Chandler Wang, Vinayak Ahluwalia, Xun Huan, Jayapalli Rajiv Bapuraj, Nikola Banovic, and Arvind Rao. 2021. Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation. *Frontiers in Neuroscience* 15 (2021), 1236. https://doi.org/10.3389/fnins.2021.740353
[42] Divya Ramesh, Vaishnav Kameswaran, Ding Wang, and Nithya Sambasivan. 2022. How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1917–1928. https://doi.org/10.1145/3531146.3533237
[43] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven (Eds.). Springer International Publishing, Cham, 19–36. https://doi.org/10.1007/978-3-319-98131-4_2
[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778
[45] Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L. Glassman, and Finale Doshi-Velez. 2021. Evaluating the Interpretability of Generative Models by Interactive Reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 80, 15 pages.

https://doi.org/10.1145/3411764.3445296

[46] Johannes Schneider, Joshua Handali, Michalis Vlachos, and Christian Meske. 2020. Deceptive AI Explanations: Creation and Detection. *CoRR* abs/2001.07641 (2020). arXiv:2001.07641 https://arxiv.org/abs/2001.07641

[47] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[48] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[49] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[50] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, et al. 2016. Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence. (2016).

[51] Simone Stumpf, Adrian Bussone, and Dympna O'sullivan. 2016. Explanations considered harmful? user interactions with machine learning systems. In *The CHI 2016 Human Centred Machine Learning Workshop*.

[52] Steven C. Sutherland, Casper Harteveld, and Michael E. Young. 2015. The Role of Environmental Predictability and Costs in Relying on Automation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2535–2544. https://doi.org/10.1145/2702123.2702609

[53] 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 *HRI '20: ACM/IEEE International Conference on Human-Robot Interaction, Cambridge, United Kingdom, March 23-26, 2020*, Tony Belpaeme, James Young, Hatice Gunes, and Laurel D. Riek (Eds.). ACM, 3–12. https://doi.org/10.1145/3319502.3374793

[54] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence* (2020).

[55] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (oct 2021), 39 pages. https://doi.org/10.1145/3476068

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

[57] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[58] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 408–416. https://doi.org/10.1145/2909824.3020230

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

[60] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI 2017, Limassol, Cyprus, March 13-16, 2017*, George A. Papadopoulos, Tsvi Kuflik, Fang Chen, Carlos Duarte, and Wai-Tat Fu (Eds.). ACM, 307–317. https://doi.org/10.1145/3025171.3025219

[61] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3025171.3025219

[62] 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, Article 2 (jan 2017), 20 pages. https://doi.org/10.1145/2998572

[63] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns (N Y)* 3, 4 (Feb. 2022). https://doi.org/10.1016/j.patter.2022.100455

[64] Enhao Zhang and Nikola Banovic. 2021. Method for Exploring Generative Adversarial Networks (GANs) via Automatically Generated Image Galleries. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 76, 15 pages. https://doi.org/10.1145/3411764.3445714

[65] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852