It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task

Patricia K. Kahr
p.k.kahr@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Gerrit Rooks
Eindhoven University of Technology
Eindhoven, The Netherlands
g.rooks@tue.nl

Martijn C. Willemsen
Eindhoven University of Technology
5600 MB Eindhoven, The Netherlands
Jheronimus Academy of Data Science
5211 DA 's-Hertogenbosch, The Netherlands
m.c.willemsen@tue.nl

Chris C. P. Snijders
Eindhoven University of Technology
Eindhoven, The Netherlands
c.c.p.snijders@tue.nl

ABSTRACT

Humans increasingly interact with AI systems, and successful interactions rely on individuals trusting such systems (when appropriate). Considering that trust is fragile and often cannot be restored quickly, we focus on how trust develops over time in a human-AI-interaction scenario. In a 2x2 between-subject experiment, we test how model accuracy (high vs. low) and type of explanation (human-like vs. not) affect trust in AI over time. We study a complex decision-making task in which individuals estimate jail time for 20 criminal law cases with AI advice. Results show that trust is significantly higher for high-accuracy models. Also, behavioral trust does not decline, and subjective trust even increases significantly with high accuracy. Human-like explanations did not generally affect trust but boosted trust in high-accuracy models.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **HCI theory, concepts and models**.

KEYWORDS

Human-AI Interaction, Trustworthy AI, Trust Development, Collaborative Decision-Making

ACM Reference Format:
Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. 2023. It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581641.3584058

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03.
https://doi.org/10.1145/3581641.3584058

1 INTRODUCTION

It might not be the exception but the rule that in only a few years mixed teams consisting of humans and artificially intelligent (AI) systems will interact with each other on a daily basis. However, collaboration with AI is not yet widely accepted, and how to best shape the interaction between humans and AI systems is not clear. The aforementioned must be confident that the AI system can help with the decision despite having a limited understanding of the AI's underlying mechanisms. Such confidence may require specific knowledge or training to become accustomed to a system. Trust in AI can be challenged even more when individuals experience failures of AI systems. One way to increase trust in model-based AI decision-making is to open the *black box* in a way that individuals feel empowered to work with it and develop a well-calibrated trust relationship. That is, they should not over-trust a failing system (for instance, because of automation or positivity bias), nor should they under-trust recommendations (for instance, because of tech skepticism, status quo bias, or general overconfidence in their own ability compared to the system).

Research has recognized the importance of knowing what influences human-AI interaction. Earlier studies regarding trust in AI have focused on the influence of system properties (e.g., XAI performance), characteristics of users (e.g., skills, attitudes), the interaction of the latter with intelligent systems, or the overall context of the decision-making task. Despite the volume and variety of existing research, trust-in-automation studies' results are inconclusive and leave lots of open issues. A large part of the trust in AI research is conducted in laboratory studies with rather abstract tasks, which offer limited options for testing complex decision-making processes, making it difficult to draw conclusions for real-world applications. In addition, few studies have investigated human-AI interaction tasks over a more extended period of time, cf. [29, 73]. However, research in trust over time is essential to understanding and developing healthy long-term relationships between humans and AI.

With the underlying study, we follow up on previous work that has examined how different explanation types and performance levels affect trust [47, 58]. We do so by analyzing trust on the basis of an applied scenario (the assessment of jail times), emphasizing

trust development over time. We consider two factors in our experiment. First, we distinguish between abstract and human-like explanations. We argue that the latter adds value in the form of a contextual, reasonable explanation that helps individuals to trust AI decisions more when compared with displaying a list of keywords. Furthermore, we consider whether trust develops differently for different model accuracies: our rationale is that AI models that perform poorly will not be able to build up the same level of trust as high-performing models. With this setup, we not only hope to contribute to the current HCI literature but can address two practical challenges simultaneously: first, measuring trust in an applied decision-making scenario by letting individuals deal with jail time decisions of real criminal law cases. Second, measuring the trust based on repeated interactions. The following research questions guide our work:

- **RQ1**: Is trust in AI different for different model accuracies and explanation types?
- **RQ2**: How does trust develop over time depending on model accuracy and type of explanation?

2 RELATED WORK

Trust in intelligent systems is a prerequisite for successful human-computer interactions. It is defined as "the attitude that an [AI] agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [44]. In many scenarios, AI advice has been shown to lead to better outcomes [48, 49, 60]. This, however, is neither a consistent finding nor does it always lead to outcomes that are as good as they can be. The aim is for humans to calibrate trust in AI correctly: they should neither place too much trust in a failing system nor put too little trust in systems for the wrong reason. Thus, we have to find ways to avoid biases in behavior, for example, blindly following AI (automation bias, positivity bias) or rejecting it on principle.

The scientific community has done considerable work to uncover processes and factors that influence trust in AI systems. Trust can be affected by the properties of the system: Transparency [10, 13, 14, 52], high system accuracy or reliability [3, 21, 25, 51, 72] increase trust, and also visual representations or human-like appearing systems [36, 37]. In addition, communication between system and humans affects trust, for example, explaining mistakes [22], repair strategies like AI providing justifications or apologies [23, 42], or written explanations - versus visuals - [68] increase AI trust. The same accounts for "human in the loop" interaction approaches [1, 18, 27]. Research also found evidence that individuals' skills and character traits influence trust: for example, confident people are found to be less trustworthy regarding AI [45], experts also tend to under-trust AI performance [49], and (political) conservatism and age are associated with low comfort of trusting AI systems [11]. Lastly, trust may depend on the context in which a task occurs: certain emotional states mediate trust in AI [24], and uncertainty and complexity hinder trust in AI [20].

2.1 Trust in AI over time

Research on trust in AI appears to be gaining recognition in terms of quantity and diversity of research themes. What still lacks attention is research of human-AI interactions in a long-term or repeated-scenario context [32, 67]. Measuring trust in AI over time adds additional complexity as we need to consider the factors just described (system properties, user characteristics, interaction modes, context) and the dynamics in the relationship between humans and AI systems. As individuals familiarize themselves with a system and understand it better, trust will grow but not necessarily in a linear fashion. Currently, research offers no clear answer as to whether trust in AI tends to decrease or increase over time. Yang et al. [71] found that trust increases with automation successes but decreases even more when seeing it err, thus summarizing that automation failures loom larger than their successes. Similar studies from Chacon et al. [12] and Nourani et al. [54] show a sharp decline in trust after early AI errors, and it was not recovering to the same levels when experiencing good AI performance afterward. Nourani and colleagues as well as Desai et al. [19] conclude a primary-recency effect: initial (and late) interactions affected trust the most when trust is measured after exposure to the system.

At the same time, some comparable studies show that trust grows over time: Chiou et al. [15] found that individuals learn to trust an intelligent (robotic) system as operators understand how to work with it over the course of several interactions successfully. Similarly, Manchon et al. [50] assessed trust in an automated driving system over time (three assessments in 4 months). In contrast to earlier studies [41], trust increased over time for both trustful and distrustful drivers. Manchon and colleagues posit that several positive interactions in the early phase of the study supported even distrustful participants to gain trust fast. It can be agreed upon that initial trust is crucial for longer interactions, especially motivating individuals to start using automated systems in the first place [16, 50]. Tolmeijer et al. [67] add that initial trust is important to accept better later mistakes, strengthening long-term interactions. Still, caution is warranted because not everyone can judge AI advice appropriately. Laypeople over-relied on AI recommendations when experiencing correct outputs at the beginning of an interaction. In contrast, experts used AI advice less, even after their performance decreased over time [54]. Beggiato and Krems [5] conclude that initial information about a system is appreciated and helps people to trust and accept AI advice. Being able to form a mental model about a system's capabilities must match with the experiences individuals make, otherwise, trust and acceptance can decrease.

Lastly, Yu et al. [73] studied trust dynamics based on different levels of model accuracy. Based on participants simulating a binary quality control task (assessing whether drinking glasses were produced correctly or not, with false positive or false negative AI advice), they suggested that trust trajectories are different based on accuracy levels: trust increased over time when system accuracy was 80% or higher. However, it decreased with 70% accuracy. They also showed that AI failures at different time stages along the interaction demonstrated different implications in the change of trust: participants that had time to become familiar with the system did not decrease in trust. Trust over time stabilized, especially at the end of each task block. Over time, individuals formed a stable mental model for themselves, also described as the inertia of trust. Overall, they propose a phase model in which individuals learn to what extent a system can be trusted (phase 1), adjust this learned trust (phase 2), and then fine-tune procedures (phase 3).

Taken together, previous research suggests that the initial phase in human-AI interactions dictates how trust in a system is developing. It is when users still get familiar with a system and form their mental model that stabilizes only after some interaction with a system. Trust increases with positive first impressions but does not recover easily from early mistakes. However, this scheme could lead to unwarranted mistrust in a system: Early failures are not always indicative of overall poor performance, and late failures do not lose severity just because people have already consolidated their own image of an AI. We now turn to how model accuracy and explanations influence trust in AI.

2.2 Trust in AI: System Accuracy

In many (though not all) prediction tasks, intelligent systems can match or outperform their human counterparts [28, 30]. In general, a system's accuracy is an essential criterion for finding a helpful AI tool and is a prominent determinant of trust [64]. Despite its capabilities, systems can fail at certain points, and users must remain alert to detect faulty systems and wrong decisions from (in principle) decent systems. Studies show that consistently low accuracy is indeed observed and acted upon: Yin et al. [72] found that trust was significantly affected by the actual level of accuracy of a system; in comparison, the effect of stated accuracy only has a negligible impact on trust. Similarly, information about a system's (high) reliability increased trust and performance [25]. Moreover, trust was sustainably damaged and recovered only slowly after experiencing errors in performance [21, 22, 51, 73]. Yu et al. [73] manipulated system accuracy on four levels (70%, 80%, 90%, and 100%), giving false positive/negative advice accordingly: trust decreased only in the 70 percent condition. Papenmeier et al. [58] compared trust levels of participants that interacted with a high, medium, and low ("antagonistic") accuracy and found that participants indeed showed adequate levels of trust, in line with Yu and colleagues. To summarize, individuals can distinguish between appropriate and poor AI advice based on its accuracy, which is a crucial prerequisite for optimal trust calibration. It is vital that individuals only trust sufficiently accurate models (that is, more accurate than they are themselves). We will apply different levels of accuracy to compare how trust develops. Like Yu or Papenmeier, model accuracy is not explicitly stated and can, therefore, only be anticipated by the recommendation itself. It will be interesting to see whether study participants can distinguish between higher and lower-performing models, and, more so, to what extent this is reflected in their trust trajectories over the course of repeated tasks.

2.3 Trust in AI: Explaining AI Output

With the increase in computing power, AI systems have also increased in complexity and opacity. The general consensus is that explanations, specifically explainable AI (XAI), enable individuals to understand AI systems and their operations better [2, 47], which is especially true for decision-making in complex, risky domains. Explanations help people to understand how, when, and why models make predictions [34]. XAI can be in textual or visual form, to complement a recommendation (e.g. counterfactual examples, probability values), as interventions (warnings, nudges), or post-decision arguments (apologies, promises, justifications) [8, 9, 22, 23, 37, 40, 55, 63, 68]. Explanations (in comparison to no explanations) have also been shown to be a success in increasing trust over time [32, 52].

Although the majority of studies found positive effects of AI explanations on trust, some studies observed feelings of manipulation by AI [9] or subsequent overconfidence in AI [69]. Papenmeier et al. [57, 58] find evidence that not all explanations are helpful, and some might even be harmful: they discovered that adding nonsensical or random explanations hurt trust (as one would hope). Furthermore, their results show that explanations do not improve trust when individuals interact with a sufficiently accurate system. They argue that the type of explanations that were used, highlighting words in a text document, did not improve decision-making as it did not add any value for participants: explanations were of statistical rather than causal nature, which only partially supports human understanding. Lim and Dey [46] found that understanding and trust in a system are highest when explanations are provided. Recent studies attempt to adapt human knowledge and rationales for XAI to increase understanding and, thus, trust in AI [43] [66]. Finally, Nourani et al. [53] tested whether trust differs for meaningful versus meaningless explanations. They defined meaningfulness as the level to which explanations were perceived as meaningful in the human context. Their results show that participants significantly underestimated system accuracy when providing weak (less human-meaningful) explanations, as they did not understood results. The study, which was conducted with non-expert participants, claims the need for "human-interpretable" explanations. As a conclusion from these research results, we hypothesize that explanations offer added value if they are human-like and contextual rather than abstract explanations that are limited in both content and form.

3 STUDY DESIGN

The underlying study builds on the knowledge of previous work: We analyze trust development using a decision-making task based on real (criminal law) data, in which participants are being supported by an AI system. Our study follows a 2x2 between-subjects design. Participants are randomly assigned to one of the four groups, in which we manipulate system accuracy (high vs. low) and explanation type (human-like vs. abstract). Their task is to estimate jail time for 20 legal cases. For this, participants are supported by an AI system that provides a numerical jail time estimation and an additional textual explanation (see Figure 1). All 20 criminal cases (from 2022) were selected from the Dutch database *de Rechtspraak* [17] and thus the final verdicts (our baseline truth for the experiment) were known.

3.1 Measurements and Hypotheses

System Accuracy. To test whether different model accuracies resulted in different trust levels, we applied two accuracy levels: one model representing an AI model with high accuracy and one model with low accuracy. Accuracy was defined as the extent to how close the jail time estimate from the AI system was to the actual jail time (ground-truth). As we wanted the deviation error to occur as natural as possible, we calculated the AI estimates by adding

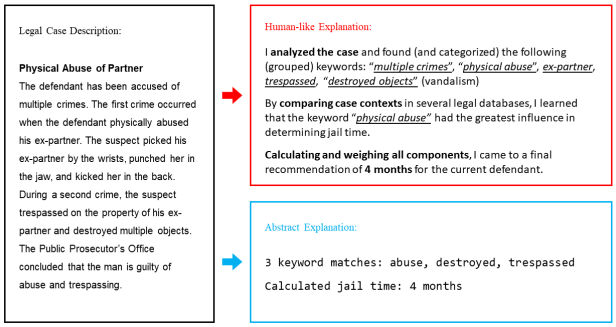

Figure 1: Exemplary Legal Case with (human-like and abstract) Explanations

a normal-distributed error around the actual jail time: The high-accuracy model deviated +/- 10 percent (95% interval) from the correct jail time, and +/- 50 percent (95% interval) in the low-accuracy model. All estimates were weighed against each other to avoid systematic bias, such that AI estimates were either higher or lower than actual jail time. Overall, we wanted participants to perceive the high-accurate model as more competent as the low-accurate model. Still, there were a few cases where the low-accuracy model performed equally well or better than the high-accuracy model (see Table 1) due to chance and the normal distribution. Unlike Yu et al. [73], we applied continuous measurements (estimates of jail time in months) and only manipulated two different accuracy levels. Finally, estimates were pre-calculated and fixed to be the same for all participants (Table 1).

Explanation Type. We expect that trust develops differently with different explanation types of a system. To prepare explanations for each case, we identified a set of case-specific keywords. For this, we first read the law case reports from which we created the shorter case descriptions. Based on the latter, we picked keywords that best represented the cases in terms of their specific analysability. The selection of case-specific keywords was then applied to two different styles of explanations: human-like explanations (condition 1) which applied full-sentence text and, thus, added a sense of human-like intelligence by reasoning more contextually, while for the second style, the abstract explanations (condition 2), the selected keywords were presented as a simple series of words. Following the AI design proposals of Knijenburg and Willemsen [38] and Nourani et al.'s [53] findings, we argue that displaying full-sentenced explanations elicits comprehensibility and allows the feeling of interacting with a rather intelligent, human-like system. Following this line of thought, we highlighted the different (explanation) capabilities also by introducing the two systems to the study participants with different system names (human-like: *AI legal case analysis system*, abstract: *jail time calculator program*) and providing details to the inner workings of each. Both, the explanations and the supporting cues as described above could lead individuals to perceive the human-like system as more capable and potentially reliable, thus enhancing trust in its recommendations.

We expect that trust is different for our four conditions: We hypothesize that participants will place more trust in a high-accuracy model versus a low-accuracy one. We furthermore believe that trust increases more (or decreases less) when explanations have a human-like portrayal. We, therefore, assume the following hypotheses, which were also pre-registered under the Open Science Framework[1].

- **H1**: Trust is higher in the high accuracy model than in the low accuracy model.
- **H2**: Trust increases more / decreases less over time in the high accuracy model than in the low accuracy model ("high accuracy protects trust better").
- **H3**: Trust is higher with human-like explanations than with abstract explanations.
- **H4**: Trust increases more / decreases less over time with human-like explanations than with abstract explanations ("Human-like explanations protect trust better").

Trust. Measuring trust development based on the presented hypotheses, we apply a two-fold approach: First, we measure behavioral trust, which Lee and See [44] define as an attitude that is directly observable and more prone to show objective outcomes. Specifically, we measure participants' Weight on Advice (WoA), which tells us how much an individual relies on the AI's advice. Sniezek and van Swol's [65] *Judge-Advisor System* paradigm calculates the degree to which people change their behavior – it is calculated by weighing two consecutive estimations, the first estimate before seeing the estimate of the automated system, and the second estimate after seeing the estimate of the automated system. The continuous outcome ranges from 0 (people completely ignored the AI estimate and stayed with their own estimate) to 1 (people completely relied on the AI advice and adopted it as their own estimate). Applying this measurement, we follow other studies in the trust in AI literature [7, 49].

$$Weight\ on\ Advice = \frac{2nd\ Estimate - 1st\ Estimate}{AI\ Advice - 1st\ Estimate}$$

As a second measurement, we track self-reported trust. Subjective measures are used to capture inherently subjective trust data, reflecting an individual's perspective [61]. These psychological constructs are usually measured with survey scales. We ask participants to indicate their level of trust after every legal case. This is done with a single-question item asking participants to indicate their current level of trust in the AI system (1 = no trust at all, 10 = full trust). Scharowski and colleagues point out that one reason the overall results of the AI literature on trust are inconclusive is that there are no standardized measures, partly because definitions of trust are ambiguous. In addition, most studies measure trust either in a self-reported or behavioral manner. To be able to compare both trust types, we use both metrics. They are measuring behavioral trust as the advantage to quantify trust somewhat objectively. However, a potential pitfall could be that not being close to the AI estimate does not necessarily mean that trust is low but that individuals are more confident in solving a task on their own. Since confidence in behavior is calculated, the result also depends on the result itself: if the user's initial estimate and the AI estimate are close from the

[1] https://osf.io/avux5/?view_only=69230d1656b14f8aaf9a6f34030bef7b

Table 1: Overview of jail time per case (from left to right): the actual jail time of each legal case, the calculated jail time estimate in the high-accuracy condition (+/- 10%), the calculated jail time estimate in the low-accuracy condition (+/- 50%)

Legal Case	Actual Jail time	High Accuracy	Low Accuracy
Stabbing in the street	24	21	10
Murder of ex-girlfriend	96	102	24
Possession of illegal fireworks	6	7	1
Import of cocaine	10	10	13
Theft in retirement home	6	5	8
Money laundering and drug possession	30	32	17
Online marketplace and Facebook Scam	4	5	4
Violence against police offer	12	12	12
Attempted murder of brother	72	81	30
Death of child	36	43	56
Theft, assault and possession of drugs	6	5	11
Attempted murder	48	52	85
Home burglary	7	7	10
Attempted murder and arson	84	75	17
Premeditated murder	120	134	126
Fraud	12	12	12
Possession of weapons	12	13	23
Physical abuse of partner	5	4	3
Gun possession and money laundering	30	28	22
Possession of weapons and cocaine	48	50	29

beginning, the deviation is naturally small. Measuring trust as a self-reported dimension can help capture an immediate psychological response based on direct experiences with AI. The drawback here is that it may depend on a personal inclination to trust AI, which is often participant-specific and cannot reflect a general attitude. Another reason for measuring trust in the explained manner is that we want to see how trust develops over time, a dimension that is often neglected. To not overburden people with too many question items, we do not use a multi-item trust checklist as recommended by Jian et al. [33] that can capture several trust dimensions (i.e., reliability, integrity, familiarity). We decided to ask for subjective trust with a single item immediately after every task similar to Yu et. al. [73].

In addition, we measure several other variables that we consider potential covariates. We ask participants to indicate their level of law expertise (10-point Likert scale: 1 = no expertise at all; 10 = very high). Logg et al. [49] showed that experience could lead individuals to over-trusting their own skills and, therefore, to mistrust or ignore an intelligent system's recommendations. Following the assumption that personality traits affect trust behavior, as done before [62], we propose to measure whether prosocialness influences trust in AI, we apply the Prosocial Behavioral Intentions Scale [4] (α = .81). Participants have to rate their willingness to get involved in social situations on a 7-point Likert scale (1 = I would definitely not do this; 7 = I would definitely do this). Finally, we ask study participants to indicate their affinity for technology, using the 4-item Affinity for Technology Interaction Short Scale (ATI-S, α = .87) from Wessel et al. [70]. Participants have to indicate to what extent they agree with the four statements on a 6-Point Likert scale (1 = completely disagree; 6 = completely agree).

3.2 Study Procedure

At the beginning of the experiment, participants were randomly allocated to one of the four groups. After being introduced to the terms of the experiment and accepting the consent form, participants were introduced to the automated system - either the system with human-like explanations, called *AI Legal Case Analysis System*, or the system with simple explanations, called *Basic Jail Time Calculator program*. To support active engagement with the system, we asked participants to confirm that they had read the introduction carefully by ticking the corresponding box. As a second introduction part, participants learned about the task procedure (Figure 2): Participants read the case and indicated their initial jail time estimate (1). After seeing the system calculate a result (interactive visual element) (2), they were automatically directed to the AI output, the numeric estimate, and the explanation (3). Participants then adjusted or confirmed their second estimate (4). They learned about the correct verdict of the case (5), and finally indicated their current trust in the AI system with a slider going from 1 = no trust at all until 10 = full trust (6).

The task procedure was repeated for 20 legal cases. To avoid order effects, cases were presented in random order. The last part of the study covered the following topics as a questionnaire: perceived level of intelligence of the AI system, perceived level of accuracy of the AI system, participant's relationship towards technological systems (affinity to technology), their willingness to participate in social situations (prosocialness), their level of law expertise, and their demographics (age, biological sex). The study closed with the debriefing and the remuneration of participants. Participants took approximately 20 minutes to finish.

We used identical AI estimates for all participants: Depending on the factors (accuracy, type of explanation) in the four conditions, all

estimates and explanations were pre-formulated and pre-calculated (see Table 1) without using an actual AI system. Even though this was part of the study's cover story, the intelligent systems were purely fictitious and did not perform any actual calculations.

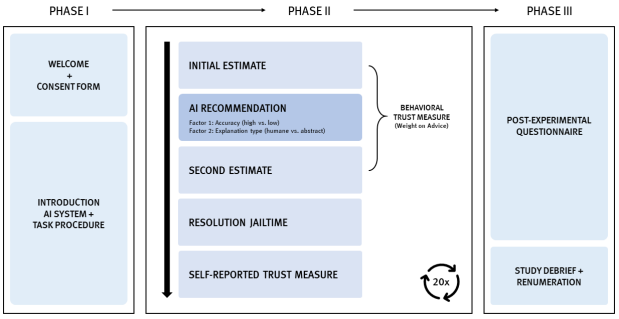

Figure 2: Experimental Study Procedure (with the main study task (Phase II) being repeated 20 times)

3.3 Participants

We calculated the sample size with an a-priori power analysis. The result was a total sample of n=171 participants, based on an ANOVA fixed effects, main effects and interactions calculation (effect size: 0.25, power: 0.9) [26]. We base our calculation on similar studies conducted by Yu et al. [73], Tolmeijer et al. [67], and Papenmeier et al. [57, 58], which reported repeated tasks with medium effect sizes (0.25).

We recruited participants via Prolific [59] based on the following features: British citizenship, aged 18 and over, with experience in the field of law (the latter characteristic only applied to half of all participants). We recruited 10 additional participants to pretest our study. Each participant received on average £5.87 as compensation. Both pre-test and study were conducted between May 24 and June 1.

4 RESULTS

204 participants registered for the study from which 171 participants (84%) fully completed it. Participants were excluded due to being under-age, not finishing all questions, or indicating unrealistic or pattern-like estimates throughout the legal task. On average, participants were 36.6 years old (Min: 19, Max: 69), 71% were female, their mean legal expertise was 4.2 (10-point Likert scale), their prosocial level average 6.0 (7-point Likert scale), and their level of tech affinity was 3.7 on average (6-point Likert scale).

Following Harvey and Fischer [31], we defined margins for behavioral trust (0 < WoA < 1). With this, we mainly excluded extreme outliers or values following uncommon behavior (e.g., always deviating away from AI advice). Trials where participants' initial estimate was identical to the AI estimate were excluded from WoA analysis as they would result in division by zero. Thus, of the 3,420 assessments made by the participants, we excluded 513 due to either WoA margins or matching estimates.

Testing whether manipulations of the conditions were recognized, we asked participants to estimate system intelligence (based on the different explanations) and system accuracy of the AI system on a 10-point Likert scale. We performed a two-sided t-test (equal variance) for both questions. Participants perceived the system with human-like explanations (M =6.39 , SD =2.45) significantly more intelligent than the system with abstract explanations (M =5.64 , SD =2.35), t (169) = −2.0 , p < 0.04. Participants also perceived the high-accuracy models (M =7.39 , SD =1.67) to be significantly more accurate than low-accuracy models (M =4.28 , SD = 1.84), t (169) = −11.55, p < 0.001. These results suggest our manipulations were successful in changing participants perception of the system.

To test our hypotheses, which were based on the effect of model accuracy and type of explanation on the development of self-reported and behavioral trust, we applied multilevel regression models with (repeated) trust measurements nested within participants. We discuss these models in detail below.

4.1 Trust is higher for high-accuracy models

To test the two hypotheses regarding the main effect of model accuracy, we analyze whether trust is higher for high-accuracy models, and whether it increases over time. We start comparing trust means by accuracy (see Table 2) and observe that for both trust measures, trust is on average higher in the high accuracy conditions: 0.61 vs 0.31, 0.49 vs 0.32 when we compare WoA for the high and low accuracy condition, and 6.80 vs 4.13 and 6.24 vs 4.14 for the self-reported trust measure (we test the statistical significance of these differences in the regression models).

We model the effects of our conditions and position of the trial (for the effect over time), as well as relevant covariates in a set of multilevel regression models on behavioral and subjective trust, as showing in Table 3. Results from regression models in Table 3 (behavioral trust, model 3: R^2 = 0.14; self-reported trust, model 6: R^2 = 0.21) support the difference in mean effects (see 2 that trust is higher for high model accuracy: both, behavioral trust (model 1: β = 0.26 , p < 0.001), and self-reported trust (model 4: β = 2.6 , p < 0.001) were significant with high accuracy (model 1 and 4 are without interaction effects of trial). In addition we see a positive interaction effect of trial and accuracy, for both trust measures. For WoA (model 4), the effect of the trial round is negative in the low accuracy condition: trust decreases over time (β = -0.005, p < 0.001). The effect of the trial round is less negative in the high accuracy condition (interaction effect β = 0.008, p < 0.001). For self-reported trust, we also observe a positive interaction of trial and accuracy (model 6). In the low accuracy condition, the effect of the trial round is zero (β = -0.003, p = 0.25), whereas it is positive in the high accuracy condition (interaction effect β = 0.057, p < 0.001). The results hence support both hypotheses H1 and H2: Trust is higher for high-accuracy models in comparison to low-accuracy models for both behavioral and self-reported measures (H1). We also find that trust is decreasing less over time (trials) when model accuracy is high compared to when it is low (H2). Figure 3 illustrates these effects graphically.

Table 2: Comparison of condition outcomes by trust measurements (95% conf. interval)

	Behavioral Trust (WoA)		Self-reported Trust	
	High Accuracy (1)	Low Accuracy (0)	High Accuracy (1)	Low Accuracy (0)
Human-like Explanations (1)	M = 0.61 SE = .01	M = 0.31 SE = .01	M = 6.80 SE = .07	M = 4.13 SE = .08
Abstract Explanations (0)	M = 0.49 SE = .02	M = 0.32 SE = .01	M = 6.24 SE = .10	M = 4.14 SE = .08

4.2 Trust is not affected by different explanation types

We moreover expected that trust will be higher for AI models with human-like explanations than for AI models with abstract (keyword) explanations (H3), and that human-like explanations protect trust better over time (H4). We first compared explanation types regarding their absolute trust levels. We find that both WoA (model 1: β = 0.05, p = 0.17) and self-reported trust (model 4: β = 0.35, p = 0.21) are not different for human-like and abstract explanations. Furthermore, we analyzed the effect of explanations over trial: in line with the previous results, there is no significant effect regarding explanation type over time on behavioral trust (interaction effect β = -0.001, p = 0.55) as well as self-reported trust (interaction effect β = -0.012, p = 0.28). Based on our findings, we reject H3 and H4: Human-like explanations do not affect trust differently than abstract explanations, they are furthermore not effectively protecting better over time based on our results.

4.3 Human-like explanations elevate trust for high-accuracy models

In addition to our hypotheses, we tested whether any interaction effects were found for model accuracy and explanation type. As it can be seen in Figure 3, comparing the lines for the two high-accuracy conditions, trust is higher with human-like explanations, esp. for behavioral trust. Our final model shows a significant interaction for behavioral trust (β = 0.123 , p = 0.03), however, this effect is not seen for self-reported trust where results are non-significant (β = 0.54 , p = 0.27). We can summarize that trust is significantly different for highly accurate models with human-like explanations. In contrast to abstract keyword explanations, human-like explanations are able to boost individuals' willingness to adapt their own advice towards the AI advice (behavioral trust).

4.4 WoA and self-reported trust only correlate to a limited extent

For our experiment, we measured trust two-fold – behavioral trust was measured with regards to the actual jail time estimates and to what extent they are influenced by the AI estimate, self-reported trust was measured after every trial to be able to retrace any change of trust immediately. Based on prior findings on trust in AI systems, we learned about the possibility of both measurements producing different outcomes. Running the empty regression models for both trust measurements to find potential differences in the overall quality of trust, we find that self-reported trust can be attributed to the individual (ρ = 0.55) whereas behavioral trust is more attributed to the study task (ρ = 0.32). This is also demonstrated in Figure 3: trust trajectories are developing differently for the same task. Connecting those findings to the main effects for accuracy (H1, H2), we again find that self-reported trust is significantly different in accuracy whereas behavioral trust only shows borderline significant outcomes. A similar discrepancy but in the opposite direction shows for interaction effects of accuracy and explanations (H4). In addition, a Pearson's correlation was run to assess the relationship between both measurements. We find a moderate correlation between behavioral and self-reported trust, r (3130) = 0.305, p < 0.00. Concluding, we interpret that trust has different "qualities", and the question remains which measurement has greater power and impact regarding trusting an AI model.

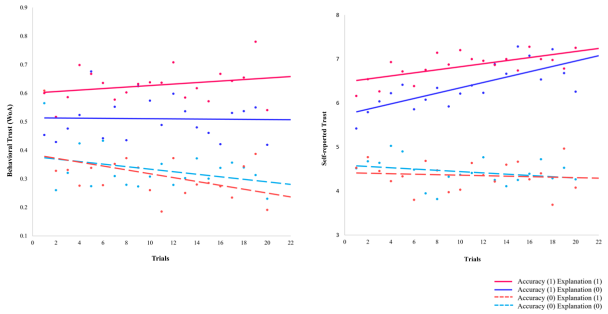

Figure 3: Development of Trust per Condition (fitted line graph)

4.5 Trust is influenced by age but not by gender, tech affinity, or legal expertise

Even though there was no focus on testing demographic or other personal characteristics of our participants, we included several covariates into one of our models (age, biological sex, legal expertise, and tech affinity) as we expected these to correlate with trust in AI systems. For example, contextual expertise was found to influence trust in AI negatively, e.g., [49], and affinity to tech promotes trust in systems [56]. We find that for both measurements trust is not significantly different for any of those variables, with one exception: age negatively affected behavioral trust (β = −0.004, p = 0.002) as seen in our final model, but self-reported trust was not (β = 0.009, p = 0.38) . This finding follows earlier arguments, for example, from Knowles and Hanson [39], that AI (and tech) aversion increases with age; as individuals become older, they grow more resistant or are not capable anymore to control, and thus, trust technology.

Besides these significant results, our overall non-significant findings are in line with research from Papenmeier et al. [58].

5 DISCUSSION

With our study, we gained new insights in (repeated) decision-making with AI support. Our results one the one hand confirms prior findings of the HCI literature and also propose new details about the interplay between model accuracy and explanation type.

Trust increases over time, but only when model accuracy is high. Our results suggest that trust is generally higher for high-accuracy models, which makes intuitive sense (even though it is not always found empirically). Moreover, trust also increases over time when participants interact with a high-accuracy model. Apparently, participants pick up the competency of the intelligent support system over time and follow its recommendation more as time progresses. However, we want to annotate that trust effects were somewhat different for our two measurements: the results for behavioral trust were less strong. A potential reason for this difference might be that if participants' initial estimate was already fairly close to the AI estimate, with - thus - no reason to deviate from the system's recommendation. This will be especially the case for cases with low jail time (e.g., 4 months) as it is more likely that the AI prediction and the participants' estimate are similar or even overlap. More testing with higher (and more homogeneous) jail times regarding the ground truth would be needed to address the assumption. Comparing our task setup with Yu et al. [73], we find similar results to their 90% accuracy condition as trust increased over time in their design as well, although we cannot compare these findings directly as the task setup was rather different. Papenmeier et al. [57] measured trust in AI for three different accuracy levels (high, medium, low) with a hate speech task and arrived at similar conclusions, adding that accuracy (vs. explanations) showed the most impact on trust overall. Their latest work [58] also confirms that high accuracy increases trust, showing once again that participants are able to pick up on an algorithm's competence, at least when it is competent enough. The question is whether there is a specific accuracy level at which trust no longer increases but decreases over time. Or, the appropriate AI accuracy level could depend on the level of the participant, where those with higher expertise need higher AI accuracy levels to be convinced.

Trust is not higher with human-like explanations. Our study reveals no significant findings of trust based on the type of explanation: there was no difference for participants that received human-like, explanations compared to participants that received abstract keyword explanations. Results were non-significant for absolute trust values and for trust development over the sequence of trials. We assumed that human-like explanations would be easier to understand and perceived as more helpful in adding value for a decision at hand, and thus are more trustworthy than just enumerating keywords. Our argumentation followed the majority of studies in the HCI literature that claims that explanations increase trust in computer models (cf. [46]). Perhaps the design of explanations were not distinct or appropriate to increase trust for the task at hand, and another design (e.g., confidence levels, visualizations [35]) could have been more helpful. Papenmeier et al. [57] conclude that not giving an explanation resulted in better or equal trust compared to giving an explanation. They reason that their approach - highlighting words in the text – might not add any trustworthy "components" to enhance understanding and, thus, increase trust in an AI system. A special case is demonstrated by the random explanation condition, which even decreases trust for the moderately and highly accurate models. We welcome the fact that random explanations do not help to build trust, as misleading measures should not be used to support decision-making after all. Even though earlier work of Papenmeier et al. [57] argues that trust was highest in the condition without any explanations, it remains to be analyzed whether trust in AI models can be improved by the right kind of explanations.

Human-like explanations boost trust for models with a high accuracy. Although different explanation types did not affect trust, we found a significant interaction effect for explanation type and model accuracy: trust was higher in high-accuracy models when combined with human-like explanations. We argue that the more sophisticated explanations corroborate the high accuracy of the model and enabled participants to understand its processes even better. Interestingly, this counter-argues findings of Papenmeier et al. [58]: in their most accurate condition, faithful explanations + high accuracy model, trust was not significantly higher. They argue that "faithful explanations do not necessarily imply meaningfulness in the eyes of the user" (p. 26) as these are based on statistical relations rather than causal information. Our findings are consistent with that individuals considered the human-like explanations supportive enough, which might be achieved by the fact that explanations followed the standard assessment processes (selecting key topics, weighing, and concluding arguments for a final estimate). Users do not blindly follow AI advice though: Human-like explanations do not increase trust in the AI estimates with low accuracy. This is in line with the results of [38].

Behavioral trust differs from self-reported trust. We measured self-reported trust and behavioral trust (WoA) after every interaction. Our results in the high accuracy condition show that self-reported trust significantly increased over the course of 20 tasks whereas behavioral trust did not. A practical reason could lie in the capability of estimating correctly: an already good initial assessment is not improved by AI advice, and respectively, weight-on-advice scores for behavioral trust would be low, regardless of the trust in the algorithm. Thus, consistent with our results, self-reported trust is high because it would be a confirmation of interacting with a helpful AI ("I am pretty much in line with the AI's estimate"), but behavioral trust would naturally be lower. Independent of the latter, these findings raise questions on the explanatory power of trust measures and the quality of trust outcomes: are individuals able to express trust adequately or is a behavioral trust measure stronger in predicting trust? Is trusting an AI as similar to relying on AI advice, cf., [6, 61].

Concluding, we have evidence to support the claim that model accuracy go with higher levels of trust (H1), and trust also increases over time (H2). We cannot conclude similarly for explanation types (H3, H4): trust is not affected, neither as absolute value nor over time. Some additional analysis showed that human-like explanations boost (self-reported) trust in high accuracy models. Therefore, we can partly answer our research question: yes, trust is affected by

Table 3: Regression Models for Behavioral and Self-Reported Trust

	Behavioral Trust (WoA)			Self-reported Trust		
	Model 1 b/se	Model 2 b/se	Model 3 b/se	Model 4 b/se	Model 5 b/se	Model 6 b/se
HumanExpl	0.048	- 0.001	- 0.007	0.195	-0.060	-0.092
	(0.03)	(0.05)	(0.05)	(0.25)	(0.38)	(0.38)
HighAcc	0.257***	0.095	0.096	2.259***	1.294**	1.313**
	(0.03)	(0.05)	(0.05)	(0.25)	(0.40)	(0.41)
Trial	-0.001	-0.004*	-0.005*	0.019***	-0.003	-0.003
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
HumanExpl_x_HighAcc		0.129*	0.123*		0.551	0.541
		(0.06)	(0.06)		(0.50)	(0.51)
Trial_x_HighAcc		0.008***	0.008***		0.057***	0.057***
		(0.00)	(0.00)		(0.01)	(0.01)
Trial_x_HumanExpl		-0.001	-0.001		-0.012	-0.012
		(0.00)	(0.00)		(0.01)	(0.01)
Age			-0.04**			0.009
			(0.00)			(0.01)
Female			-0.049			-0.123
			(0.03)			(0.28)
TechAffinity			-0.011			-0.027
			(0.02)			(0.13)
LegalExpertise			-0.008			0.047
			(0.01)			(0.05)
Constant	0.317***	0.386***	0.595***	4.072***	4.465***	3.987***
	(0.03)	(0.04)	(0.07)	(0.23)	(0.27)	(0.59)
R-Square	0.1176	0.1284	0.1379	0.1986	0.2051	0.2075

*p <0.05, ** p <0.01, *** p <0,001*

(high) model accuracy, and it positively develops over time when accuracy is high. However, human-like explanations do not generally affect trust (development) positively.

6 LIMITATIONS AND FUTURE STUDIES

Our work uncovered new insights in trust development in AI systems, shedding light on model accuracy and explanation efforts in particular. Still, we want to discuss some limitations of our work. We found that model accuracy is affecting trust over time: an AI system with high accuracy, deviating 10% away from the correct resolution versus an obviously flawed system with low accuracy, deviating up to 50%. The question remains whether the same results are obtained when the accuracy design changes. What would be the accuracy level at which participants could still learn and adjust accuracy patterns whether the system is right or wrong? Manipulations in our study could have been too obvious and, thus, interpreting trust is too simple to that extent. Yu et al. [73] applied four different levels of accuracy, which resulted into different trust trajectories. Future work could test different or more accuracy levels at once. In addition, we challenge the definition of accuracy and its effect on trust: individuals might perceive an overall predictable (steadily running) model higher in trust than an overall highly-accurate one that lacks predictability. Is a system more trusted when giving steady support but overall lower accuracy than a system that is very high in accuracy but with the tendency to include few outliers? A final suggestion would be to analyze trust with regards to the perception of a system being able to learn and improve over time, which we would define as adaptive accuracy. Would this AI behavior increase trust as this demonstrates a natural behavior for humans as well when improving over time?

Human-like explanations did not result in higher AI trust (development) across both the low and high accuracy models. In hindsight, using an extra control condition in which no explanations are used, would have allowed additional and useful comparisons. Key to our finding is that although explanations do not work throughout, they do work for high accuracy AI models. Our study showed that users were distrustful when experiencing odd AI behavior - in our case receiving seemingly intelligent explanations for what actually was a poor recommendation. In contrast, human-like explanations led participants to trust a highly accurate AI model more than when explaining a case with only simple keywords. Obvious further lines of inquiry would be to explore which kinds of explanations work best, and which other communication efforts affect trust. One could think of semantic feedback after model mistakes (apologies, justification), or task framing (solving a task together vs. AI is support), both of which seem likely to be able to affect trust.

Table 4: Hypotheses Overview

	Hypothesis	Result
H1	Trust is higher in the high accuracy model than in the low accuracy model.	Supported
H2	Trust increases more / decreases less over time in the high accuracy model than in the low accuracy model ("high accuracy protects trust better").	Supported
H3	Trust is higher with human-like explanations than with abstract explanations	Not supported
H4	Trust increases more / decreases less over time with human-like explanations than with abstract explanations ("human-like explanations protect trust better").	Not supported

Although we have endeavored to create a plausible study task based on real legal data, and in this sense have created an experiment that is much closer to real life than in abstract lab studies, the question remains as to whether this is close enough. One could argue that the setup of the task, especially resolving each case at the end (participants got feedback on the actual jail time after each trial), is not depicting a real-life procedure, but teaches especially inexperienced participants to focus on observing (and relying on) what the AI model suggested. In addition, actual jurisprudence involves a more comprehensive, iterative approach - a shortcoming that some study participants with a legal background commented on. Besides these experimental design considerations, there are also some other implications of the task: jail sentencing standards are not only dependent on the country and the legal expert who is deciding, but also does not provide a clear objective ground truth (other than that in the presented case, the jail time happened to be of a particular magnitude). In this sense, participants had to predict in a noisy environment, where each particular case can be seen as just one example of what could happen. An alternative would be to take as ground truth the mean (or median) jail time for similar cases in the database. In addition, a participant's potential belief that AI should not be involved in such a typically human decision-making task may have affected our results, although participant level variance is relatively small in our data. We obviously acknowledge the importance of discussing various ethical issues regarding the use of AI in a legal context, but want to emphasize that they exceed the scope (and focus) of this work, which was to observe behavioral change in response to receiving AI support in a complex decision-making task.

Lastly, with our repeated task scenario, we aimed to contribute to the study of trust development. It is unclear whether 20 successive interactions in a single trial of 20-30 minutes is sufficient for this purpose. For example, Tolmeijer and colleagues [67] offer as a compromise to repeat interactions over a period of a few weeks. We hope to be able to collect longer-term observations of human-AI collaboration over several months to draw more robust conclusions about the dynamics and effects of trust over time.

7 CONCLUSION

Decision-making in various fields is becoming more and more a hybrid performance of humans and intelligent systems. There are several arguments in favor of combining human talent and AI capabilities to achieve the best results. However, there is still some guessing what we need to do to design applications trustworthy enough. With our study we were able to demonstrate that model accuracy has significant effects on trust development of individuals. We found no conclusive differences in the way an outcome was explained, which we argued would support individuals' decision-making. However, we found that human-like explanations could elevate trust when model performance was high. With our results, we were able to contribute new insights related to the development of trust in the context of a real-life task. We want to continue exploring how system performance, individuals' expectations, and additional circumstances influence trust. Our ultimate goal is to enable people to successfully calibrate their trust when systems act outside of the expected frame, similar to how humans sometimes do.

ACKNOWLEDGMENTS

We thank Luc Siecker, Jane Deijnen, Milo Simons, Lorea Ros, and Ruben van der Werf for their help in conducting the study. In addition, we would like to express our gratitude to the European Supply Chain Forum (ESCF), the department Industrial Engineering and Innovation Sciences (IE&IS), the Eindhoven Artificial Intelligence Systems Institute (EAISI), and the Logistics Community Brabant for sponsoring the research project *AI Planner of the Future*, which is funding the Ph.D. project "Trust in AI over time" and thus supporting this and future studies.

REFERENCES

[1] Naomi Aoki. 2021. The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior* 114 (2021), 106572.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[4] Rachel Baumsteiger and Jason T Siegel. 2019. Measuring prosociality: The development of a prosocial behavioral intentions scale. *Journal of personality assessment* 101, 3 (2019), 305–314.

[5] Matthias Beggiato and Josef F Krems. 2013. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour* 18 (2013), 47–57.

[6] Michaela Benk, Suzanne Tolmeijer, Florian von Wangenheim, and Andrea Ferrario. 2022. The Value of Measuring Trust in AI-A Socio-Technical System Perspective. *arXiv preprint arXiv:2204.13480* (2022).

[7] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. 2021. Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering* 63, 1 (2021), 55–68.

[8] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[9] Christopher Burr, Nello Cristianini, and James Ladyman. 2018. An analysis of the interaction between intelligent software agents and human users. *Minds and machines* 28, 4 (2018), 735–774.

[10] Christopher S Calhoun, Philip Bobko, Jennie J Gallimore, and Joseph B Lyons. 2019. Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. *Journal of Trust Research* 9, 1 (2019), 28–46.

[11] Noah Castelo and Adrian F Ward. 2021. Conservatism predicts aversion to consequential Artificial Intelligence. *Plos one* 16, 12 (2021), e0261467.

[12] Alvaro Chacon, Edgar E Kausel, and Tomas Reyes. 2022. A longitudinal approach for understanding algorithm use. *Journal of Behavioral Decision Making* (2022).

[13] Chih-Yang Chao, Tsai-Chu Chang, Hui-Chun Wu, Yong-Shun Lin, and Po-Chen Chen. 2016. The interrelationship between intelligent agents' characteristics and users' intention in a search engine by making beliefs and perceived risks mediators. *Computers in Human Behavior* 64 (2016), 117–125.

[14] Jessie YC Chen, Michael J Barnes, Anthony R Selkowitz, Kimberly Stowers, Shan G Lakhmani, and Nicholas Kasdaglis. 2016. Human-autonomy teaming and agent transparency. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*. 28–31.

[15] Manolis Chiou, Faye McCabe, Markella Grigoriou, and Rustam Stolkin. 2021. Trust, shared understanding and locus of control in mixed-initiative robotic systems. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 684–691.

[16] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.

[17] de Rechtspaak. 2022. *de Rechtspraak Website*. https://www.rechtspraak.nl/

[18] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering* 61, 5 (2019), 637–643.

[19] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 251–258.

[20] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science* 31, 10 (2020), 1302–1314.

[21] 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 (2015), 114.

[22] Mary Dzindolet, Linda Pierce, Scott Peterson, Lori Purcell, Hall Beck, and Hall Beck. 2002. The influence of feedback on automation use, misuse, and disuse. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46. SAGE Publications Sage CA: Los Angeles, CA, 551–555.

[23] Connor Esterwood and Lionel P Robert. 2021. Do you still trust me? human-robot trust repair strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 183–188.

[24] Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Theodore Jensen, Yusuf Albayram, and Emil Coman. 2021. Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. In *Designing Interactive Systems Conference 2021*. 1492–1503.

[25] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R Endsley. 2008. The influence of agent reliability on trust in human-agent collaboration. In *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*. 1–8.

[26] 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. *Behavior research methods* 41, 4 (2009), 1149–1160.

[27] Juliana Jansen Ferreira and Mateus Monteiro. 2021. The human-AI relationship in decision-making: AI explanation to support people on justifying their decisions. *arXiv preprint arXiv:2102.05460* (2021).

[28] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 1–8.

[29] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[30] William M Grove and Paul E Meehl. 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, public policy, and law* 2, 2 (1996), 293.

[31] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes* 70, 2 (1997), 117–133.

[32] 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.

[33] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[34] Uday Kamath and John Liu. 2021. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer.

[35] Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre-Majorique Léger. 2022. Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience* 16 (2022).

[36] Rabia Fatima Khan and Alistair Sutcliffe. 2014. Attractive agents are more persuasive. *International Journal of Human-Computer Interaction* 30, 2 (2014), 142–150.

[37] Taenyun Kim and Hayeon Song. 2021. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics* 61 (2021), 101595.

[38] Bart Knijnenburg and Martijn Willemsen. 2016. Inferring Capabilities of Intelligent Agents from Their External Traits. *ACM Transactions on Interactive Intelligent Systems* 6 (11 2016), 1–25. https://doi.org/10.1145/2963106

[39] Bran Knowles and Vicki L. Hanson. 2018. The Wisdom of Older Technology (Non)Users. *Commun. ACM* 61, 3 (feb 2018), 72–77. https://doi.org/10.1145/3179995

[40] Spencer C Kohn, Daniel Quinn, Richard Pak, Ewart J De Visser, and Tyler H Shaw. 2018. Trust repair strategies with self-driving vehicles: An exploratory study. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. Sage Publications Sage CA: Los Angeles, CA, 1108–1112.

[41] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied ergonomics* 66 (2018), 18–31.

[42] Maier Fenster1and Inon Zuckerman2and Sarit Kraus. 2012. Guiding user choice during discussion by silence, examples and justifications. In *ECAI 2012: 20th European Conference on Artificial Intelligence*, Vol. 242. IOS Press, 330.

[43] Samantha Krening and Karen M Feigh. 2018. Characteristics that influence perceived intelligence in AI design. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 1637–1641.

[44] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[45] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. 2000. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104.

[46] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[47] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. 2019. Why these explanations? Selecting intelligibility types for explanation goals.. In *IUI Workshops*.

[48] Tyler J Loftus, Patrick J Tighe, Amanda C Filiberto, Philip A Efron, Scott C Brakenridge, Alicia M Mohr, Parisa Rashidi, Gilbert R Upchurch, and Azra Bihorac. 2020. Artificial intelligence and surgical decision-making. *JAMA surgery* 155, 2 (2020), 148–158.

[49] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[50] JB Manchon, Mercedes Bueno, and Jordan Navarro. 2021. Calibration of Trust in Automated Driving: A Matter of Initial Level of Trust and Automated Driving Style? *Human Factors* (2021), 00187208211052804.

[51] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch. 2012. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making* 6, 1 (2012), 57–87.

[52] Marieke Möhlmann and Lior Zalmanson. 2017. Hands on the wheel: Navigating algorithmic management and Uber drivers'. In *Autonomy', in proceedings of the international conference on information systems (ICIS), Seoul South Korea*. 10–13.

[53] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.

[54] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[55] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.

[56] Atte Oksanen, Nina Savela, Rita Latikka, and Aki Koivula. 2020. Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology* 11 (2020), 568256.

[57] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[58] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction*

(TOCHI) 29, 4 (2022), 1–33.
[59] Prolific.co. 2022. *Prolific Research Platform.* https://www.prolific.co/
[60] Timothy M Rawson, Raheelah Ahmad, Christofer Toumazou, Pantelis Georgiou, and Alison H Holmes. 2019. Artificial intelligence can improve decision-making in infection management. *Nature Human Behaviour* 3, 6 (2019), 543–545.
[61] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures. *arXiv preprint arXiv:2203.12318* (2022).
[62] Navya Nishith Sharan and Daniela Maria Romano. 2020. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon* 6, 8 (2020), e04572.
[63] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.
[64] Donghee Shin, Bu Zhong, and Frank A Biocca. 2020. Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management* 52 (2020), 102061.
[65] Janet A. Sniezek and Lyn M. Van Swol. 2001. Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes* 84, 2 (2001), 288–307. https://doi.org/10.1006/obhd.2000.2926
[66] Andrea Tocchetti and Marco Brambilla. 2022. The Role of Human Knowledge in Explainable AI. *Data* 7, 7 (2022), 93.
[67] 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.
[68] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* IEEE, 109–116.
[69] Adrian Weller. 2019. Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 23–40.
[70] Daniel Wessel, Christiane Attig, and Thomas Franke. 2019. ATI-S-an Ultra-Short scale for assessing affinity for technology interaction in user studies. In *Proceedings of Mensch und Computer 2019.* 147–154.
[71] X Jessie Yang, Christopher Schemanske, and Christine Searle. 2021. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *arXiv preprint arXiv:2107.07374* (2021).
[72] 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.* 1–12.
[73] 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.* 307–317.

Received 14 October 2022; revised 3 February 2023; accepted 11 February 2023