Master's programme: Applied Cognitive Psychology (ACP)
Student: Anzhelika Sosnova
ID: 7668554
Supervisor: Dr. M. J. Mulder
Auditor: Robert Weijers
Project: Thesis 27.5 ECTS
Thesis place: Utrecht University
Period: 1.02.2023—31.06.2023

Utrecht University

Thesis report

Understanding Perception of Algorithmic Predictions

Abstract

Decision making is often supported by forecasts of different sources including human experts and artificial intelligence. This paper examines perception of algorithmic and human-made forecasts and its potential influence on associated decision making in situations of uncertainty. Two groups of participants were given the same set of hypothetical choice problems with embedded in them probabilistic forecasts. One group was presented those forecasts as made by human experts and the second group was told that the forecasts were made by AI. Every choice problem proposed an option that participants must accept or reject. The objective of this study was to observe preferability of choice options in Human and AI conditions depending on the forecasts' framing (positive or negative), confidence level (high, medium, low) and decision domain (serious or trivial). We found that in general AI-made forecasts receive less "yes" answers to the choice problems than human-made forecasts. Overall, the framing and confidence levels affected the probability of a "yes" response. However, only the framing showed different magnitude of the phenomena in AI and Human condition indicating an interaction between those two factors. No main effect was discovered for the decision domains of the questions. Additionally, a trust scale revealed higher trust levels towards a human expert when compared to trust in AI. These findings contribute to the psychology human-AI interaction and decision making under uncertainty and suggest that people see algorithmic predictions as lacking trustworthiness.

Introduction

Artificial intelligence (AI) has become an increasingly powerful tool in many fields, including healthcare (Saravi et al., 2022; Chassagnon et al., 2021; Jin et al., 2022), finance (Nosratabadi et al., 2020; Cohen, 2022), and technology, due to its growing capabilities and favourable outcomes. With progressions in AI development, the accuracy of predictions made by these algorithms has improved significantly (Braga-Neto, 2020). For example, recent advancements were made in traffic forecasting using AI technology (Shaygan et al., 2022). Furthermore, AI (including advanced statistics and Machine Learning) is proposed to be used to enhance the accuracy of predicting students' performance in institutions of higher education (Denes, 2023; Zeineddine et al., 2021) and it is being used to anticipate a heart attack (Hutson, 2017). As AI is increasingly used to make predictions that impact human lives, it becomes crucial to understand how individuals perceive and interpret these forecasts, as such perceptions can significantly impact the effectiveness and adoption of algorithmic decision support systems. For example, it will help to understand how consumers assess information given by AI and use AI technology in decision making (Longoni et al., 2019). Second, it helps create positive and meaningful experiences with technology (Mcknight, et al., 2011). Third, it contributes development of trustworthy AI and shaping user's perceptions of AI attributes (e.g., usefulness, performance, accuracy, credibility) in algorithm-driven technologies (Choung, David, Ross, 2022). This paper aims to investigate the factors influencing human perception of algorithmic forecasts and explore the implications for decision-making.

Prior research has highlighted the role of factors as transparency, explainability, trustworthiness and reliability in shaping human perceptions of algorithmic systems (Choung, David & Ross, 2022; Glikson & Woolley, 2020). In relation to human evaluation of forecasts, studies show people's desire for good *explanations* of predictions themselves, their components, and historical data that was analysed (Yates et al., 1996). And when experts generally provide a clear rationale to underpin their forecasts, the reasoning underlying a statistical forecast may be unavailable or, even if it is supplied, it may be mysterious to those not trained in statistics (Önkal et al, 2009). Thus, algorithmic forecasts lack transparency and explainability in consumers' perception. Many researchers support the idea of perceived trustworthiness of human forecasters based on their ability to explain not only the forecast itself but also errors in their forecasts by attributing them to causes other than their own expertise (Tetlock, 2005). At the same time, errors of algorithmic forecasters make them lose their reliability in the user's eyes (Dietvorst et al., 2014). On the other hand, AI is more likely to evoke a concern that recipients' unique characteristics and circumstances will be neglected (Longoni et al., 2019). *Uniqueness neglect* was indicated to be a barrier in adoption of medical AI (Longoni et al., 2019). Those combine findings suggest perceived superiority of human forecasters when compared to algorithmic forecasters.

Besides the source, there are forecast types depending on their content. Deterministic forecasts imply certainty of a future event taking place when *probabilistic* forecasts indicate that it has some specified chance of occurring (Nadav-Greenberg & Joslyn, 2009). Given a choice people prefer deterministic forecasts over probabilistic ones (Tetlock & Gardner; 2016) likely because it makes assessment and decision-making easier. However, it was proven by several experimenters that people can take advantage of uncertainty information to make better decisions, as compared with the same decisions made without uncertainty information (Nadav-Greenberg & Joslyn, 2009; Grounds, Joslyn, Otsuka, 2017; Joslyn & Savelli, 2010). Prior studies on judgement under uncertainty have discovered that subjective evaluation of probabilities diverge from the formal probability calculus (Kahneman & Tversky, 1973; Weber, 1994). A probability has different decision weight depending on the value of the outcome and the desirability of the prospect (Keren & Teigen, 2001). In the light of decision-making, *extremeness* of forecasts seems to determine their usefulness (Yates et al., 1996). For instance, a 90% probability is preferred to a 70% probability and 70% is more valuable and informative than 50% (Keren & Teigen, 2001; Yates et al., 1996). Thus, the further a probability level from 50% the better it facilitates decision-making. In this case, could a forecast's extremeness outplay preferability of its source? How does it affect overall evaluation of algorithmic predictions?

Forecasts also vary by the domain they made in. Consumers may be less willing to rely of algorithms for more consequential tasks (serious domain) because doing so poses greater risk (Castelo et al., 2019). There's evidence that people trusted a human expert more than algorithms with subjective tasks (e.g., dating advice) but were willing to trust algorithms on some objective tasks like financial advice and data analysis (Castelo et al., 2019). Similarly, people viewed human decisions as fairer and more trustworthy in social tasks (e.g., predicting students' success in college) but with mechanical tasks (predicting stock prices) perceived fairness and trustworthiness were equal for humans and algorithms (Lee et al., 2018). We propose that the reason for the pattern of undervaluing algorithmic predictions in emotional or social domain might be related to uniqueness neglect mentioned before (Longoni et al., 2019). In the situations that typically involve social interaction, AI is seen as a poor substitute. However, the study that proposed uniqueness neglect as a reason for algorithm aversion tested AI acceptability only in healthcare which is a domain of emotional, social, and serious high-risk decision-making. We find it important to compare perception of algorithmic forecasts made in serious and trivial domains because perceived risk can influence one's willingness to rely on AI systems (Gulati, Sousa, Lamas, 2019).

Aside from forecasts' source and content, individuals exhibit sensitivity towards the manner in which this information is communicated — *how forecasts are framed.* Different ways of presenting the same information often evoke different emotions (Kahneman & Tversky, 1992; Kahneman, 2011). The statement that "odds of survival one month after surgery are 90%" is more *reassuring* that the equivalent statement that "mortality within one month of surgery is 10%". Similarly, yogurt that is "90% fat-free" is more *attractive* than "10% fat" yogurt. The equivalence of the alternative formulations is transparent, but an individual normally sees only one formulation when deciding to buy yogurt or agree to a surgical procedure. It was discovered that the framing effect has potential to affect decision-making in a wide range of situations as it's studied in fields of psychology (Wiseman & Levin, 1996; Gosling & Moutier, 2019), healthcare (Gong et al., 2013), microeconomics (Rahman, Crouch, Laing, 2018), sociology (Cheon et al., 2021). However, there isn't a clear comparison of the framing effect manifestation for AI-generated and human-created forecasts. Will the phenomena have the same magnitude? To what degree it will affect the decision-making? Prior research does not have a clear answer to these questions. The present study intended to fill this gap.

To test for a possible perceptual bias against AI-generated forecasts in different decision domains, for different probability levels, and framing options, we set up a mixed design experiment. Using examples of previous research on decision-making under uncertainty, we created a set of twenty-four choice problems. Each problem was phrased in a positive or negative frame, had a prediction in one of the confidence levels (high, medium, low) and belonged to a serious or trivial domain. Those problems formed two identical sets (24 questions in each): one had AI as prediction-maker and the second had human-experts as forecasters. The two surveys were distributed in two groups of participants, which for the rest of this paper we call Human condition and AI condition. Additionally, we measured the levels of perceived trust to see if the results of the choice problems survey corresponded with level of trust towards a forecaster. The trust scale for Human condition had questions about trustworthiness of medical professional and the trust scale for AI condition consisted of questions about trustworthiness of medical AI. For this experiment we adopted human computer trust scale (by Gulati et al., 2019). Overall, we expected lower preferability of choice options in AI condition and lower trust towards AI compared to human experts. On the other hand, we expected to see high preference towards high probability forecasts, and that acceptance of algorithmic forecasts is significantly higher in the trivial domain (as compared to algorithmic forecasts in the serious domain). Furthermore, we explored the framing effect occurrence in Human and AI conditions.

Methods

A mixed-design online experiment was conducted in May–June 2023.

Participants

We recruited 106 (mean (std) age = 26.3 (7.3), 78 female, 26 male, 2 non-binary) participants from Utrecht University and the surrounding area to fill an online survey that took around 15 minutes to complete. Participants had to be at least 18 years-old and be able to read and comprehend English. If a participant was an Utrecht University student, they were granted participation points.

For Human condition, a total of 48 completed responses were obtained. We omitted participants who filled in the survey in less than five minutes ($N = 5,\ 10\%$) as we tested that less than five minutes was too little time to read and understand every question. Out of remaining 43 participants 33 were female (77%) and 10 were male (23%). For AI condition, a total of 49 valid responses were obtained. The analysis excluded participants who filled the survey in under five minutes ($N = 9,\ 18\%$). Out of remaining 40 participants 28 were female (70%), 11 were male (28%), and 1 participant non-binary (3%). See the combined overview of the demographic characteristics of the participants in Table 1.

	Gender	
Male	21	25%
Female	61	73%
Non-binary	1	1%
Total	83	
	Age	
18–24	51	61%
25–34	17	20%
35–49	13	16%
50+	2	2%
Mean	26,5	
Median	23	
Mode	22	
	Familiarity with AI technology	
Daily	10	12%
Occasionally	70	84%
Never	3	4%
	Native language	
Dutch	60	72%
English	6	7%
Romanian	3	4%
Chinese	3	4%
Other	10	12%

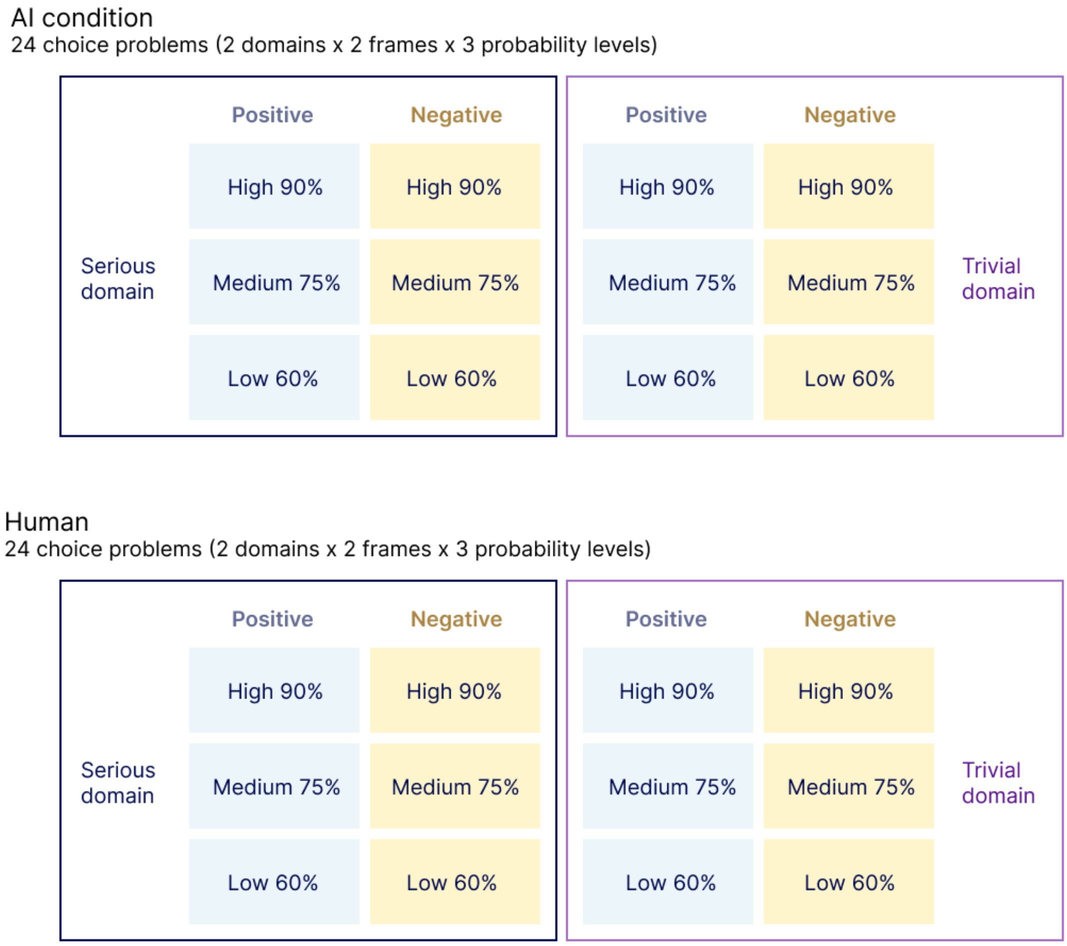
Table 1. Summary of study participants by gender, age, native language, and familiarity with AI technology

Electronic informed consent was obtained from all the participants, through Quatrics. This study involving human subject research received full approval by the Ethical Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University on 17 May 2023 (ref. no 23-1624).

Design and Materials

The core of our experiment was a set of hypothetical choice questions. The hypothetical problems similar to ours were frequently used in psychology of decision making (Kahneman & Tversky, 1984; Johnson et al., 1993) and proved themselves reliable for testing human perception (Kühberger, Schulte-Mecklenbeck, & Perner, 2002; Wiseman & Levin, 1996). Two surveys were created, one for each condition (AI and Human). A set of hypothetical choice problems consisted of twenty-four statements with a mandatory choice in each of them: accept or reject a proposed option.

A 2 (positive versus negative frame) × 3 (high, medium, low probability level) × 2 (serious versus trivial question domain) withing participants — with addition of human versus AI between participants factor — mixed design was used. Each participant filled in a survey with twenty-four choice problems. Among the twenty-four problems twelve were created in the serious domain. Those were mostly medical questions that had to make participants think about life endangering situations (e.g., serious illnesses and risky treatments). The other twelve choice problems were made in the trivial domain to make people think about mundane everyday decision (e.g., choice of a movie or a restaurant). Each domain question was phrased either in a positive or negative frame and had an imbedded probability of one of the three levels (high, medium, low). For the forecasts' confidence levels, we have chosen the probabilities of 90, 75, and 60 per cent. Thus, each domain included six of positively framed and six of negatively framed choice problems. And each of those frame groups included two questions with a high probability forecast, two with a medium probability forecast and two with a low probability forecast. So, all our factors were equally represented in the survey (see Figure 1). For example, the question "On a summer day, an AI algorithm predicted 40% chance of hail and 60% chance no hail. Would you plan outdoor activities?" belongs to AI condition, framed negatively, has a low confidence forecast and is part of trivial domain questions. The content of the questions for AI and Human conditions matched exactly except for a forecast maker.



AI condition
24 choice problems (2 domains x 2 frames x 3 probability levels)

Human
24 choice problems (2 domains x 2 frames x 3 probability levels)

Figure 1. Hypothetical choice problems design. Each condition has a set of 24 choice problems configured in the serious or trivial domain, a positive or negative frame and one of the probability levels: high 90%, medium 75%, low 60%. 2 questions in each probability level make up to 6 questions per frame (in one of the domains) and 12 questions per domain.

Trust Scale

To explore people's trust in AI forecasting capabilities and compare it to the degree of trust towards human forecasters, we included a simple trust scale at the end of the survey for both conditions. We adopted a validated trust scale (HCTS) except for the questions that wouldn't be suitable for measuring trust in a human expert (Gulati et al., 2019). For example, "I feel I must be cautious when using (—)" measures perception of risk in human-computer interaction but loses its usefulness within interpersonal trust context. The refined scale consisted of seven questions that included measurements of benevolence, competence, and reliability (Table 2). As the authors of the scale, we used a 5-point Likert scale — where 1 indicates strongly disagree and 5 indicates strongly agree — to collect participants responses.

Trust component	Corresponding item
Benevolence	I believe that a doctor/medical AI will act in my best interest.
Benevolence	I believe that a doctor/medical AI is interested in understanding my needs and preferences.
Competence	I think that a doctor/medical AI is competent and effective in making predictions about treatment outcomes.
Competence	I think that a doctor/medical AI performs its role as a forecast maker very well.
Competence	I believe that a doctor/medical AI has all the qualifications I would expect from a forecaster.
Reliability	I think I can completely depend on the prognosis made by a doctor/medical AI.
Reliability	I can trust the information presented to me by a doctor/medical AI.

Table 2. HCTS questions

Procedure

Participants were randomly assigned to one of the two conditions: AI or Human. After given consent they answered basic demographic questions. Next, in AI condition subjects were presented with twenty-four choice problems formulated as "an AI algorithm predicted…" with the following details of a forecast and an option that an individual was offered to accept or decline. In Human condition, the twenty-four choice problems had identical information in them, and they were formulated with a human forecaster, for example: "A scientist predicted effectiveness of a new medicine as 90% chance it will cure the disease and 10% it will weaken the immune system. Would you take the medicine?" In the middle of the choice problems block a comprehension question was asked (both conditions) to ensure that participants pay close attention to the task at hand. After the choice problems block participants were asked to fill in the human-computer trust scale (HCTS).

In each trial the order of the choice problems was automatically randomised to eliminate potential influence of the order on the experiment's results. The completion time was recorder for each participant, and it averaged 13.6 minutes.

Analysis

After cleaning (discarding invalid responses) and formatting (text into numerical values) the data, we constructed several data visualizations to look for patterns and compare key-characteristics in it (Figure 2 & 3). We calculated the probability of a "yes" answer per participant and then compared group mean values to see the preferability of positively and negatively framed choices as well as the preferability of forecasts with high, medium, and low confidence levels.

To understand how different factors (and their interaction) predict the probability of saying "yes" to a choice problem, a logistic regression was performed on the data sets of AI and Human condition groups. We employed a generalized linear mixed model (GLMM) to account for potential interaction effects when a "yes" answer in a one of the conditions might depend on (1) the forecast confidence level or (2) the questions' domain, (3) the framing of the option. We used a GLMM with choice options preferability as the response variable, condition group, forecasts' confidence level, questions' domain, and questions' framing as the fixed effects variables, and the participants ID as the random effect to capture within-subject correlation. The logit link function and binomial error distribution were selected to model the binary nature of choice options preferability data (yes = 1, no = 0). The GLMM showed good fit to the data, as indicated by the log-likelihood value (−1086.15) and the AIC (2376.29).

Furthermore, Pared Samples *t*-Test was used to test significance of the framing effect in AI and Human conditions. We used the results to compare the strength of the framing effect in the two conditions (Figure 3).

Additionally, Independent Samples *t*-Test was used to compare trust levels (as well as trust components: benevolence, competence, reliability) between the two condition groups (AI and Human). See Figure 6 & 7.

Results

The Choice Problems

To quantify preferability of different forecasts' confidence levels, we calculated the probability of a "yes" answer per participant for the high, medium, and low confidence forecasts. Then we calculated group mean probabilities per confidence level per condition (Table 5).

Condition	Forecast	Mean *p* of a "yes" answer	SD
AI	90%	0.76	0.18
AI	75%	0.51	0.21
AI	60%	0.48	0.18
Human	90%	0.84	0.13
Human	75%	0.64	0.17

Human	60%	0.52	0.15

Table 5. Probability (group averages) of an "yes" answer for high, medium, and low predictions' confidence level per condition.

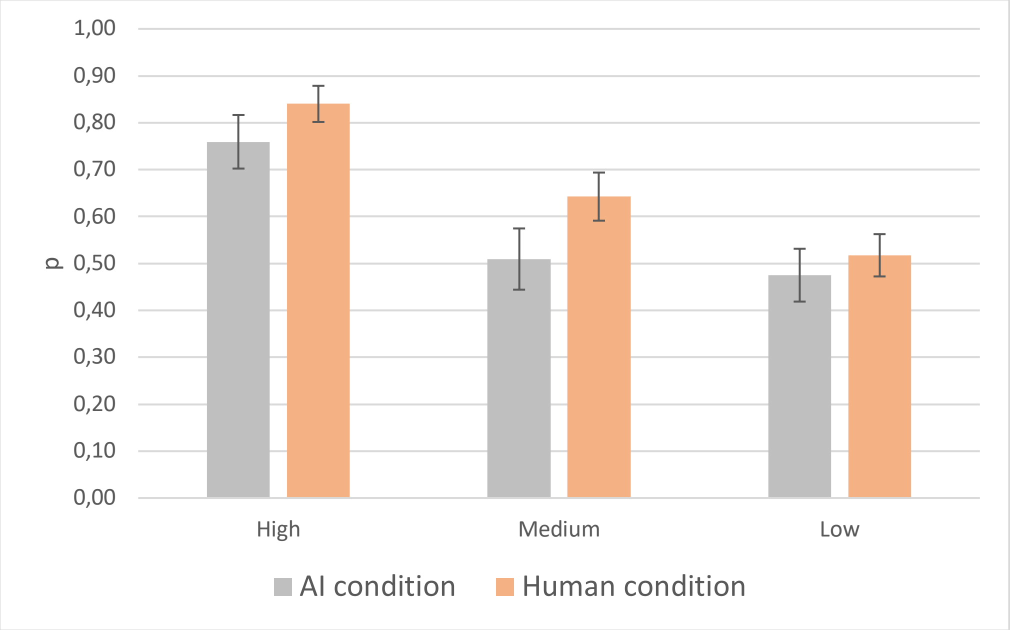


Figure 2. A comparison of group means for probability of a "yes" answer: high, medium, and low probability forecasts in two conditions. Error bars represent 95% confidence intervals. AI = artificial intelligence.

The same way as for the confidence levels, we calculated the probability of a "yes" answer per participant for negatively and positively framed questions. From those values we calculated group mean probabilities of a "yes" answer for positive and negative frames in AI and human condition (Table 6).

Condition	Frame	Mean *p* of a "yes" answer	SD
AI	Positive	0.62	0.17
AI	Negative	0.55	0.16
Human	Positive	0.74	0.15
Human	Negative	0.59	0.13

Table 6. Mean group probability of an "yes" answer for positive and negative frames per condition.

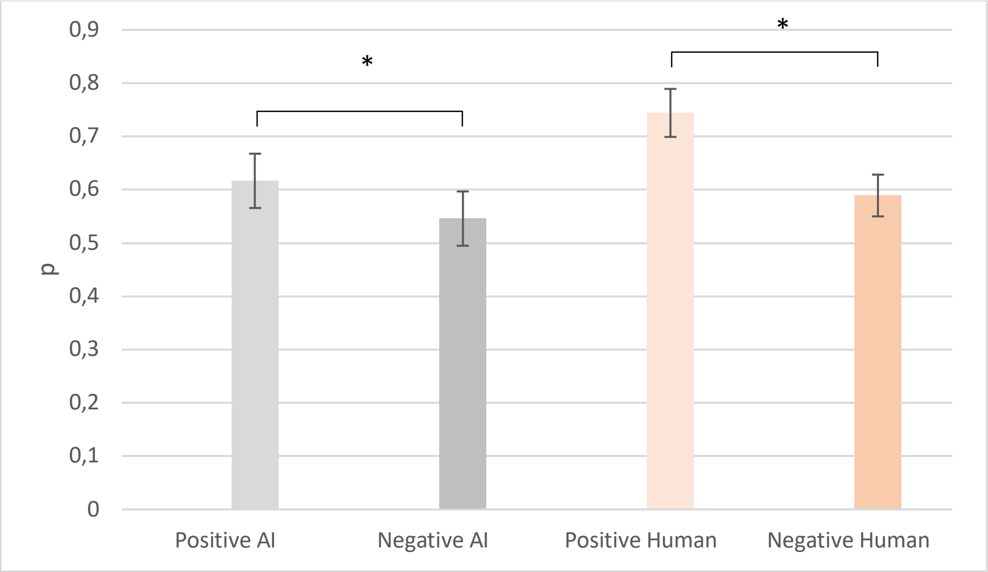


Figure 3. A comparison of group means for probability of a “yes” answer: positive and negative frames in two conditions. Error bars represent 95% confidence intervals. AI = artificial intelligence. Paired Samples *t*-Test showed a significant framing effect in *AI condition* scores between positive frames $(M = 0.62, SD = 0.17)$ and negative frames $(M = 0.55, SD = 0.16)$, $t\ (39) = 2.47$, $p = 0.018$. A significant framing effect was found in *Human condition* scores between positive frames $(M = 0.74, SD = 0.15)$ and negative frames $(M = 0.59, SD = 0.13)$, $t\ (42) = 6.04$, $p < 0.001$.

To understand how the condition, forecasts’ confidence level, decision domain, and the framing of the options predict the probability of saying “yes” to a choice problem, a logistic regression has been used. We found a main effect for condition $(X^2(2) = 10.86, p < .001)$ suggesting that participants in the two condition groups had significantly different choice preferability. The probability of a “yes” answer to a choice problem was significantly higher in Human condition than in AI condition. This finding indicates a difference in perception of probabilistic forecasts based on a forecaster (AI or human). Contrary to our expectation no main effect was found for the questions’ domain $(X^2(1) = 1.07, p = 0.3)$. See more details in Table 7. Moreover, framing $(X^2(1) = 18.79, p < .001)$ and confidence level $(X^2(2) = 70.15, p < .001)$ showed significant main effects. However, the model yielded an interaction effect of condition and framing $(X^2(1) = 6.45, p = 0.011)$. The interaction effect indicates a higher likelihood of a “yes” answer to a *positively framed* choice problem in Human condition when compared to AI condition. Paired Samples *t*-Test revealed a significant framing effect in both conditions. The results showed a significant difference between mean probabilities of accepting a choice option in AI condition between positive frames $(M = 0.62, SD = 0.17)$ and negative frames $(M = 0.55, SD = 0.16)$, $t\ (39) = 2.47$, $p = 0.018$. And a significant framing effect was found in Human condition probability scores between positive frames $(M = 0.74, SD = 0.15)$ and negative frames $(M = 0.59, SD = 0.13)$, $t\ (42) = 6.04$, $p <$

0.001. Thus, even though we observed the framing effect in both conditions its *intensity* was greater in Human condition.

Effect	df	X^2	*p*
Confidence	2	70.147	< .001
Framing	1	18.794	< .001
Domain	1	1.073	0.300
Condition	1	10.855	< .001
Confidence ✻ Framing	2	0.930	0.628
Confidence ✻ Domain	2	49.022	< .001
Framing ✻ Domain	1	0.000	1.000
Confidence ✻ Condition	2	0.077	0.962
Framing ✻ Condition	1	6.449	0.011
Domain ✻ Condition	1	0.000	1.000
Confidence ✻ Framing ✻ Domain	2	9.749	0.008
Confidence ✻ Framing ✻ Condition	2	0.217	0.897
Confidence ✻ Domain ✻ Condition	2	3.993	0.136
Framing ✻ Domain ✻ Condition	1	0.000	1.000
Confidence ✻ Framing ✻ Domain ✻ Condition	2	0.145	0.930

Table 7. Generalized Linear Mixed Model with logit link function. ANOVA Summary

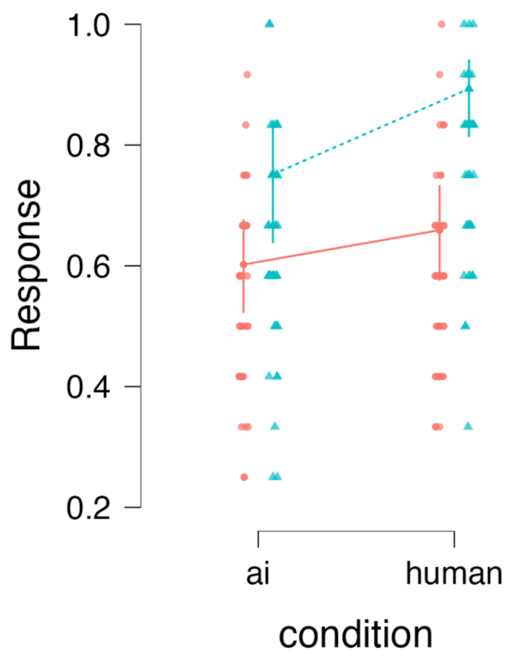


Figure 4. Probability of a "yes" answer in the two condition groups. Red represents negative frames and teal blue represents positive frames. Error bars represent 95% confidence intervals. AI = artificial intelligence.

The logistic regression results showed two more interaction effects: the interaction effect between the forecasts' confidence level and domain (serious or trivial), and a three-factors interaction among the confidence level, framing and domain. Those effects were left out of the analysis because the condition doesn't play a role in them, hence it falls outside of the focus of this study.

Trust in a Human Expert versus Trust in AI Forecaster

Descriptive HCTS data. The 5-point Likert scale was converted into numerical values, from 1 point for "strongly disagree" to 5 points for "strongly agree". Those values per participant — AI and Human conditions compared — are visualised in a box-plot bellow (Figure 5). Formulating the questions we used examples from the medical realm — asked about trusting beliefs in a doctor for Human condition and medical AI for AI condition. However, because of broad nature of the HCTS questions and the fact the participants filled the questionnaire after the choice problems set, we believe that the results reflect the general trust levels in a human expert or AI.

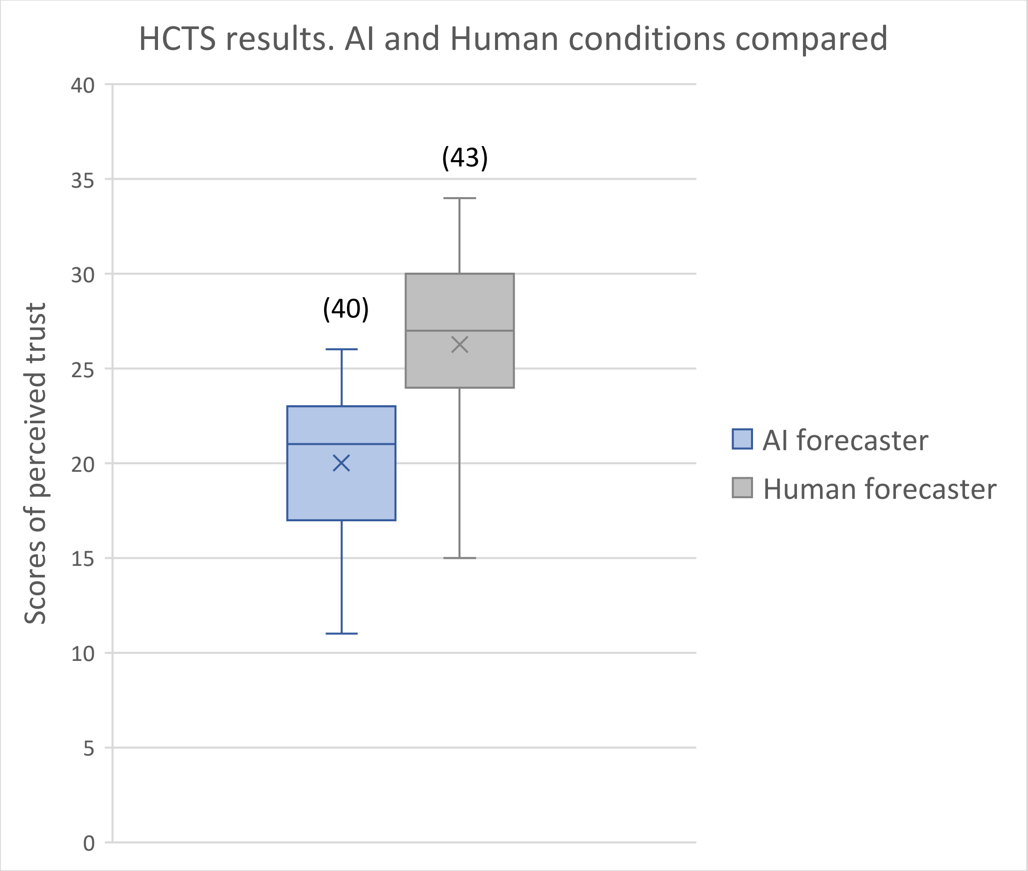


Figure 5. The distribution of HCTS data. Values in brackets are the number of participants. For each group, whiskers represent minimum and maximum scores; the box represents the interquartile range (IQR), the medial (the line in the middle) and the mean (the x) values.

To determine if there is a difference in perceived trust for AI and human forecasters we run Independent Samples t-Test. The results showed a significant difference between the mean trust scores for AI ($M = 20.00, SD = 4.07$) and a human forecaster ($M = 26.28, SD = 4.87$), $t\ (81) = -6.35$, $p < 0.001$. This result suggest that perceived trust is considerably higher in a human expert than in AI capabilities.

	N	Mean	SD	SE
Trust AI	40	20.0	4.1	0.6
Trust Human	43	26.3	4.9	0.7

Table 3. Independent samples t-Test. Student's t-Test. Descriptive statistics

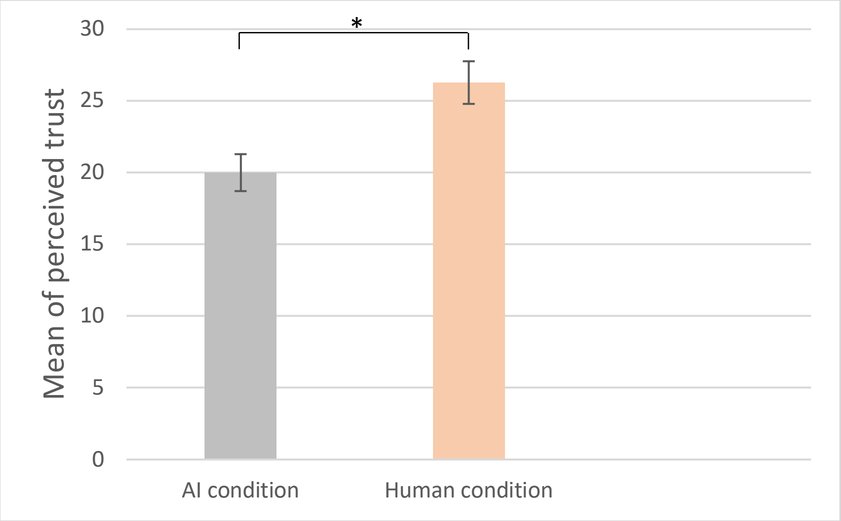


Figure 6. Trust scores means by conditions. Error bars represent 95% confidence intervals. AI = artificial intelligence. $p < 0.001$

HCTS measures three trust attributes: benevolence, competence, and reliability (Gulati et al., 2019). To see if there's significant difference in every of the trust attributes, we conducted Independent Samples t-Test for the trust components' scores. A significant effect was found in *benevolence* scores between the AI forecaster ($M = 6.18, SD = 1.86$) and the human forecaster ($M = 7.70, SD = 1.82$), $t\ (81) = -3.76$, $p < 0.001$. A significant effect was found in *competence* scores between the AI forecaster ($M = 9.28, SD = 2.14$) and the human forecaster ($M = 11.16, SD = 2.18$), $t\ (81) = -3.98$, $p < 0.001$. A significant effect was found in *reliability* scores between the AI forecaster ($M = 4.55, SD = 1.20$) and the human forecaster ($M = 7.42, SD = 1.78$), $t\ (81) = -8.56$, $p < 0.001$. Figure 7 displays a comparison of group mean scores for the three trust components in AI and Human conditions.

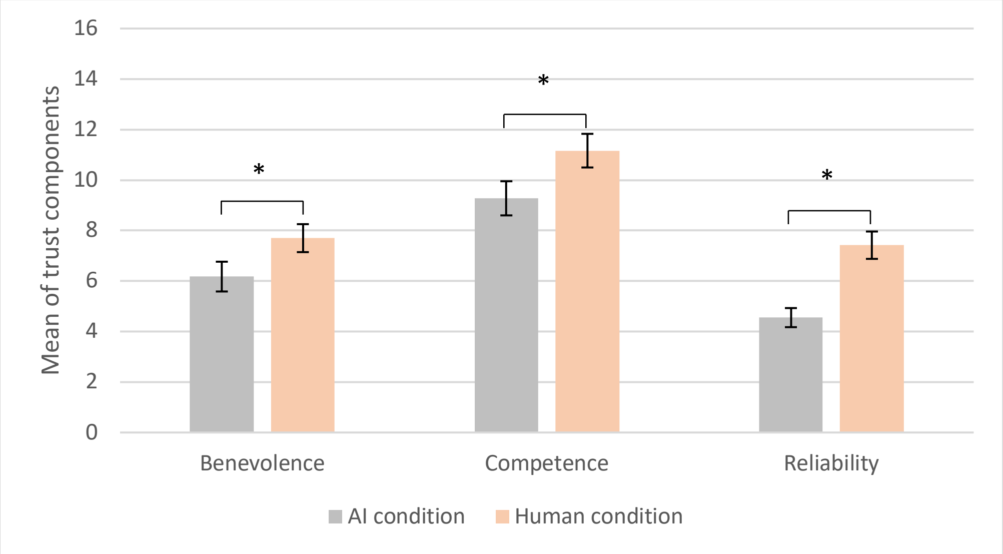


Figure 7. Trust components means by conditions. Error bars represent 95% confidence intervals. AI = artificial intelligence. $p < 0.001$

These results showed that perceived trust towards a human expert is higher compared to trust in AI (using an example from the medical realm where trust in medical professionals is compared to trust in medical AI).

Discussion

The current study examines perceptions of probabilistic forecasts made by AI and cognitive biases that play a role in human judgement under uncertainty. We compared preferability of positively and negatively framed choice options (that were created in the serious and trivial domains and included high, medium, and low probabilities) in the contexts of AI-generated and human-made forecasts to test if there is a difference in perceptions.

The main finding of our experiment was that people's perceptions of human-made forecasts is significantly more favourable than of AI-made forecasts and that perceived trust levels are higher for human experts compared to AI. These results support previous findings that people have more confidence in their own or other human expert forecasting abilities rather than in AI's predictive capabilities (Dietvorst et al., 2014), and that forecasting advice was discounted more if it came from a statistical model rather than a human expert (Önkal et al., 2009). In line with prior studies, our experiment demonstrated human's sensitivity to a forecast's source (human or AI) and lower self-reported trust levels were towards AI than towards human experts. In real-life scenarios, a preference for human advice may be explained by the opportunity to request an *explanation* of a forecast from a human advisor (Yates et al., 1996), or the notion that a human expert will better accommodate a person's *unique needs* and circumstances (Longoni et al., 2019). This study has shown that, even in a hypothetical situation — as compared to a face-to-face conversation with a human expert — people still prefer a forecast from a human over an algorithmic prediction.

Choices of hight, medium and low probabilities showed a main effect. The finding is consistent with the literature on consumers' assessment of probabilities (Teigen, 2001; Yates et al., 1996). It is worth noting that people's appreciation for probability's extremeness outshone the bias against algorithmic forecasts. The preferability of high confidence predictions in AI condition was still higher than preferability of medium confidence predictions in Human condition. That discovery might help us in the future to mitigate the perception barrier for AI-made forecasts.

Contrary to our expectations, the decision domain didn't show significant influence on options preferability. That means that choice problems from serious and trivial domains were treated by the participants the same way in both conditions. We expected that trivial domain choices with implied low risk and light to negligible consequences would help people use AI-generated forecasts and rely on them. No influence of the decision domain might be related to the fact that AI scored lower in competence and reliability levels. Thus, the influence of perceived competence and reliability of AI is strong enough to be dominant in both (the serious and trivial) domains.

Moreover, our experiment provided comparison of the framing effect manifestation for AI and human-made forecasts. Significant differences in choice options preferability (positively framed over negatively framed) were found in both conditions. However, the framing effect was more prominent in Human condition than in AI condition. And an interaction effect was found between framing and condition. We propose that participants in AI condition were more prone to choosing a "no" answer and that reduced the framing effect visibility. That means in various practical settings that we must account for the framing effect not to provoke overreliance on positively framed forecasts and under reliance on negatively framed forecasts, besides accounting for the confirmed by the present study bias against algorithmic predictions.

In addition to the hypothetical choice problems, the participants filled in a trust survey which demonstrated higher trust levels towards human experts compared to AI. We found statistically significant differences for combined trust levels between conditions as well as for every trust component: benevolence, competence, and reliability. This finding supports the main outcome of the research that perception of human forecasters in more favourable than algorithmic forecasters and it is adding to the literature on trust in automation (Cannizzaro, 2020; Longoni, Bonezzi, Morewedge, 2019; Yokoi et al., 2021). Our research points to a critical barrier that consumer-facing AI-based forecasting will need to overcome to gain acceptance and trust: consumers might see predictions as not reliable on the grounds that they come from AI. Changing this belief will be fundamental to utilise the full potential of AI-made forecasts to benefit our society in the future.

Despite the solid evidence of the phenomenon collected during our study, we see limitations that offer several opportunities for future research. First, we recognise that the sample size and the demographic characteristics of our participants weren't diverse enough to be population representative. It's reported by previous research, that younger adults (age 18–28) are quicker to adapt new technology and they are more frequent users of technology as well (Olson et al., 2011). Thus, younger adults potentially more accustom using AI technology and willing to trust and rely on it more than other age groups. When the experiment is re-created with balanced age representation in sample groups, the bias against algorithmic predictions might strengthen or weaken. Second, there's research proving that gender plays an important role in shaping individual technology adoption and usage (Morris & Ackerman, 2000). Studies show that males use new technology more frequently than females presumably due to males moving through the adoption process more quickly than females (Li, Glass, Records, 2008). That insight advocates for gender-balanced experiment groups at least in the field of new technology adoption. Aside from age and gender, there're other demographic characteristics that may influence trust in technology among people. For example, educational background and occupation. Therefore, we acknowledge the limitations of our experiment execution and hope that it can be re-done in the future with a larger, more heterogeneous sample groups.

Conclusion

Algorithmic forecasts are increasingly used in many industries making it important to better understand human perception of algorithmic forecasts as compared to forecasts made by human experts. When offered a choice problem with imbedded prediction made by *a human expert*, people are more incline to accept the proposed option than when the same forecast was created by *AI* (preferability of algorithmic forecasts seems to be less dependent on how a prediction was framed). Special attention is paid to confidence levels of predictions as high probabilities are preferred over medium and low probabilities. Phrasing of forecasts influence subjects' preferability as well, they are more likely to accept a positively framed option (and reject a negatively framed one). At the same time, the decision domain (serious or trivial) has no significant influence on the way people treat choice problems. In addition, perceived trust levels are higher towards human experts than towards AI. The findings from this research have important practical implications for developers of algorithmic forecasters and decision support systems. In particular, the results shed light on a general barrier in working with and relying on AI that not only lay people but experts in forecasting must overcome to appropriately use AI technology. We propose that individuals' belief in superiority of human experts' judgment may be motivated by desire to get a good explanation (Yates et al., 1996) or to be accommodated in one's unique needs (Longoni et al., 2019). Whatever the cause, the task of educating people to give an equal or greater weight to algorithmic predictions (rather than human-made) is likely to be a difficult one.

References

Braga-Neto, U. (2020). Fundamentals of pattern recognition and machine learning. *Springer Nature Switzerland*. https://doi.org/10.1007/978-3-030-27656-0

Cannizzaro, S., Procter, R., Ma, S., Maple, C. (2020). Trust in the smart home: Findings from a nationally representative survey in the UK. *PLoS ONE* 15(5). https://doi.org/10.1371/journal.pone.0231615

Castelo, N., Bos, M. W., Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research:* Vol. 56, No. 5, pp. 809–825. https://doi.org/10.1177/0022243719851788

Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T. N., Dangeard, S., Deutsch, E., Andre, F., Guillo, E., Halm, N., El Hajj, S., Bompard, F., Neveu, S., Hani, C., Saab, I., Campredon, A., Koulakian, H., Bennani, S., Freche, G., Barat, M., Lombard, A., Fournier, L., Monnier, H., Grand, T., Gregory, J., Nguyen, Y., Khalil, A., Mahdjoub, E., Brillet, P. Y., Tran Ba, S., Bousson, V., Mekki, A., Carlier, R. Y., Revel, M. P., Paragios, N. (2021). AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. *Medical Image Analysis* https://doi.org/10.1016/j.media.2020.101860

Cheon, J. E., Nam, Y., Kim, K. J., Lee, H. I., Park, H. G., Kim, Y.–H. (2021). Cultural variability in the attribute framing effect. *Frontiers in Psychology.* https://doi.org/10.3389/fpsyg.2021.754265

Choung, H., David, P., Ross, A. (2022). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*: Vol. 39, No. 9 https://doi.org/10.1080/10447318.2022.2050543

Cohen, G. (2022). Algorithmic trading and financial forecasting using advanced artificial intelligence methodologies. *Mathematics,* Vol. 10. https://doi.org/10.3390/math10183302

Denes, G. (2023). A case study of using AI for General Certificate of Secondary Education (GCSE) grade prediction in a selective independent school in England. *Computers and Education: Artificial Intelligence* https://doi.org/10.1016/j.caeai.2023.100129

Dietvorst, B. J., Simmons, J. P., Massey C. (2014). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology* http://dx.doi.org/10.1037/xge0000033.supp

Glikson, E., Woolley, A. W. (2020). Human trust in Artificial Intelligence: Review of empirical research. *Academy of Management Annals:* Vol. 14, No. 2, 627–660. https://doi.org/10.5465/annals.2018.0057

Gong, J., Zhang, Y., Yang, Z., Huang, Y., Feng, J., Zhang, W. (2013). The framing effect in medical decision-making: a review of the literature. *Psychology, Health & Medicine:* Vol. 18, No. 6, pp. 645–653 http://dx.doi.org/10.1080/13548506.2013.766352

Gosling, C. J., Moutier, S. (2018). Is the framing effect a framing affect? *Quarterly Journal of Experimental Psychology:* Vol. 72, No. 6, pp. 1412–1421. https://doi.org/10.1177/1747021818796016

Grounds, M. A., Joslyn, S., Otsuka, K. (2017). Probabilistic interval forecasts: An individual differences approach to understanding forecast communication. *Advances in Meteorology:* Vol. 2017. https://doi.org/10.1155/2017/3932565

Gulati, S., Sousa, S., Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology,* 38:10, 1004–1015. https://doi.org/10.1080/0144929X.2019.1656779

Ho, S. M., Ocasio-Velázquez, M., Booth, C. (2017). Trust or consequences? Causal effects of perceived risk and subjective norms on cloud technology adoption. *Computers & Security,* Vol. 70, pp. 581-595. http://dx.doi.org/10.1016/j.cose.2017.08.004

Hutson, M. (2017). Self-taught artificial intelligence beats doctors at predicting heart attacks. *Science Magazine,* April 14. https://dx.doi.org/10.1126/science.aal1058

Jin, W., Dong, S., Yu, C., Luo, Q. (2022). A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning. *Computers in Biology and Medicine* https://doi-org.proxy.library.uu.nl/10.1016/j.compbiomed.2022.105560

Johnson, E. J., Hershey, J., Meszaros, J., Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty,* 7, 35–51.

Joslyn, S., Savelli, S. (2010). Communicating forecast uncertainty: public perception of weather forecast uncertainty. *Meteorological Applications,* Vol. 17, pp. 180–195.

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux.*

Kahneman, D., Tversky, A. (1992). Advances in Prospect Theory. *Journal of Risk and Uncertainty,* Vol. 5, pp. 297–323.

Kahneman, D., Tversky, A. (1984). Choices, values, and frames. *American Psychologist,* 39:4, pp. 341–50.

Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica,* Vol. 47, No. 2.

Kahneman, D., Tversky, A. (1973). On the psychology of prediction. *Psychological Review,* Vol. 80, pp. 237–251.

Kühberger, A., Schulte-Mecklenbeck, M., Perner, J. (2002). Framing decisions: Hypothetical and real. *Organizational Behavior and Human Decision Processes,* 89, 1162–1175.

Lankton, K. N., McKnight, D. H., Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems:* Vol. 16, No. 10, pp. 880–918.

Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors:* Vol. 46, No. 1, pp. 50–80.

Lee, M. K., Rich, K. (2021). Who is included in human perception of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. *CHI Conference on Human Factors in Computing Systems* https://doi.org/10.1145/3411764.3445570

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society.*

Levin, I. P., Schneider, S. L., Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes:* Vol. 76, No. 2, pp. 149–188.

Li, S., Glass, R., Records, H. (2008). The influence of gender on new technology adoption and use — mobile commerce. *Journal of Internet Commerce:* Vol. 7, No. 2, pp. 270–289. https://doi.org/10.1080/15332860802067748

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

Mcknight, D. H., M. Carter, J. B. Thatcher, P. F. Clay. (2011). Trust in a Specific Technology: An Investigation of Its Components and Measures. *ACM Transactions on Management Information Systems* Vol. 2, No. 2: 12. http://doi.acm.org/10.1145/1985347.1985353

Mellers, B. A., Lu, L., McCoy, J. P. (2023). Predicting the future with humans and AI. *Society for Consumer Psychology:* Vol. 6, No. 1. https://doi.org/10.1002/arcp.1089

Morris, M. G., Ackerman, P. L. (2000). A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. *Organizational Behavior and Human Decision Processes:* Vol. 83, No. 1, pp. 33–60.

Nadav-Greenberg, L., Joslyn, S. L. (2009). Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making,* Vol. 3, No. 3, pp. 209–227.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., Reuter, U., Gama, J., Gandomi, A. H. (2020). Data science in economics: Comprehensive review of advanced machine learning and deep learning models. *Mathematics,* Vol. 8. http://dx.doi.org/10.3390/math8101799

Olson, K. E., O'Brien, M. A., Rogers, W. A., Charness, N. (2011). Diffusion of technology: Frequency of use for younger and older adults. *Ageing International:* Vol. 36, No. 1, pp. 123–45.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., Pollock, A. (2009). The relative influence of advice from human expert and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*: Vol. 22, pp. 390–409.

Raghavan, M., Barocas, S., Kleinberg, J., Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Conference on Fairness, Accountability, and Transparency,* https://doi.org/10.1145/3351095.3372828

Rahman, A., Crouch, G. I., Laing, J. H. (2018). Tourists' temporal booking decisions: A study of the effect of contextual framing. *Tourism Management:* Vol. 65, pp. 55–68.

Saravi, B., Hassel, F., Ülkümen, S., Zink, A., Shavlokhova, V., Couillard-Despres, S., Boeker, M., Obid, P., and Lang, G. M. (2022). Artificial Intelligence-Driven Prediction Modeling and Decision Making in Spine Surgery Using Hybrid Machine Learning Models. *Multidisciplinary Digital Publishing Institute* https://doi.org/10.3390/jpm12040509

Shaygan, M., Meese, C., Li, W., Zhao, X., Nejad, M. (2022). Traffic prediction using artificial intelligence: Review of recent advances and emerging opportunities. *Transportation Research Part C: Emerging Technologies* https://doi.org/10.1016/j.trc.2022.103921

Tetlock, P. E., Gardner, D. (2016). Superforecasting. The Art and Science of Prediction. *Random House Books.*

Tversky, A., Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *Journal of Business,* 59:4, 5251–78.

Tversky, A., Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review,* Vol. 102, No. 2.

Yates, J. F., Price, P. C., Lee, J. W., Ramirez, J. (1996). Good probabilistic forecasters: The 'consumer's' perspective. *International Journal of Forecasting.*

Yokoi, R., Eguchi, Y., Fujita, T., Nakayachi, K. (2021). Artificial intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity. *International Journal of Human-Computer Interaction,* Vol. 37, No. 10, pp. 981–990. https://doi.org/10.1080/10447318.2020.1861763

Weber, E. U. (1994). From Subjective Probabilities to Decision Weights: The Effect of Asymmetric Loss Functions on the Evaluation of Uncertain Outcomes and Events. *Psychological Bulletin.*

Wiseman, D. B., Levin, I., P. (1996). Comparing risky decision making under conditions of real and hypothetical consequences. *Organizational Behavior and Human Decision Processes,* Vol. 66, No. 3, pp. 241–250.

Zeineddine, H., Braendle, U., Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers and Electrical Engineering,* 89. https://doi.org/10.1016/j.compeleceng.2020.106903