Trust in Human-bot Teaming: Applications of the Judge Advisor System

Jonathon Love, Quentin Gronau, Scott D. Brown, and **Ami Eidels**

School of Psychological Science, University of Newcastle, Australia
Callaghan, NSW 2308 AUS

Abstract

Recent years have seen remarkable advances in the development and use of Artificial Intelligence (AI) in image classification, driving cars, and writing scientific articles. Although AI can outperform humans in many tasks, there remain domains where humans and AI working together can outperform either working alone. For humans and AI to work together effectively, the human must trust the AI bot to the right degree (calibrated). If the human does not trust the bot sufficiently, or conversely trusts the bot more than is warranted, the human-bot team will not perform as well as they could. We report three experiments examining trust in human-AI teaming. While existing studies typically collect binary responses (to trust, or not to trust), we present a novel paradigm that quantifies trust in a bot-recommendation in a continuous fashion. These data allow better precision, and in the future the development of more refined models of human-bot trust.

Keywords:

Judge Advisor System; trust; Human-bot teaming; AI

Introduction

The last decade has seen remarkable advances in the development and use of Artificial Intelligence (AI) across a range of fields. AI has been able to perform faster and more reliably than humans in a number of fields, leading to them replacing human decision-making in these areas.

Although AI can outperform humans in many tasks, there remain areas where AI either does not perform as well as humans, or where AI performance is different from humans – that is, humans outperform AI in some portion of the task, and AI outperforms humans in some parts of the task. For example, Tejeda, Kumar, Smyth, and Steyvers (2022) demonstrated an image classification task where humans outperform AI on some images, and AI outperforms humans on others.

In these instances, humans and bots working together can achieve better performance than either could on their own, be it in accuracy or speed or both. For this reason, there is growing interest in "human-bot teaming" for domains where high-quality decisions are important.

A crucial component to this human-bot teamwork is the exercise of trust. If the human fails to trust the bot sufficiently, they may ignore or not sufficiently incorporate an accurate bot's recommendation into their final decision. Similarly, if the human comes to over-trust the bot, relying on it greater than is warranted, then once again, the human-bot team will fail to achieve optimal performance. Trust must therefore be appropriately "calibrated" (Lee & Moray, 1994; Muir, 1987).

Lee and See (2004) defined human-machine trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p 51). If the human trusts the machine, they are more likely to make use of it. In this way, trust is a latent psychological construct, and is typically measured through self-report (Razin & Feigh, 2023).

To explore human-bot trust, Bansal et al. (2019) had participants classify items, based on a bot recommendation, as either defective or not necessarily defective. Each of the items varied on a number of attributes (colour, shape, size), and the accuracy of the bot's recommendations varied based on these attributes. Bansal et al. (2019) characterized this as an "error boundary", a boundary separating the attribute combinations that the bot classifies correctly, from the attribute combinations that the bot classfies incorrectly. Bansal et al. found that when the "error boundary" was simple (could be represented with few points), and non-stochastic (each recommendation from beyond the error boundary is always wrong), people developed more accurate mental models of the error boundary and achieved higher accuracy.

Similarly, Yu et al. (2017) presented participants with a fictitious computerised task, where they were to classify glasses as either "broken" or "not broken", and could receive a recommendation from a bot. Participants demonstrated some level of calibration in their trust: where the recommendations were more accurate, participants were more likely to rely on them.

These studies have practical value in demonstrating the conditions in which bot recommendations can improve team's performance, but from a measurement perspective they only provide a very coarse measure of trusting behaviour (which then limits the ability to develop and test theoretical models). The measure is the proportion of times the participant decides to adopt, or rely on, the bot's recommendation. However, it may be that the participant does not actually trust the bot, but rather has no reason to distrust the bot. It is also possible that the participant neither trusts nor distrusts the bot, but when forced to decide between the two, they elect to trust. In this way, what might look to be very high trusting behaviour, say, relying on the bot in over 90% of trials, may actually represent a much lower level of trust.

In the present study we developed a more precise measure of trust, that provides a finer-grained trust score, and in the

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

future will allow better traction for shared mental models of human and bots.

Overview of the experiments

Experiment 1 provides a conceptual replication of Yu et al.'s (2017) study, with an additional condition in which the participant is offered a (new) third option; "don't know". We expect participants to make use of this response when they neither trust nor distrust the recommendation. This procedure provides a somewhat coarse measure of human trust in the bot's recommendation. In Experiments 2 and 3 we extract a continuous and more precise measure of trust using a variation of the Judge Advisor System (JAS).

Experiment 1

Experiment 1 is a replication of Yu et al. (2017)'s study with an additional condition allowing the participant to respond "don't know".

Methods

In this task, participants play the role of a worker in a light-bulb factory whose job it is to determine if each light-bulb coming off a production line is broken or not. To assist with this process, a helpful cartoon robot provides a recommendation as to whether the light-bulb should be classified as broken or not. The participant then provides their decision by selecting from buttons labelled "broken" or "not broken". Following this, the true state of the light-bulb is revealed, and participants receive a fictitious $100 reward if they are correct, or fictitious $100 penalty if they are incorrect. Significantly, the participant only had the robot's advice to inform their decision. There is no information apart from the robot's recommendation on which the participant could base their decision. The second condition varied from the first in that it provided the participant a third button, a "don't know" response, which allowed them to opt out of making a decision one way or the other.

Each participant completed 5 blocks of 20 trials. Each block provided the participant with a different helpful robot (coloured differently, and introduced at the beginning of the block with a different name), with varying levels of accuracy. The five accuracy levels were 100% (the robot's recommendation was always correct), 90%, 80%, 70% and 60%. In the first block the participant always received 100% accurate recommendations, with the remaining blocks in random order.

8 participants were recruited on the Prolific experimental platform, with 4 assigned to the replication condition, and 4 assigned to the "don't know" condition. Participants completed the task in around 20 minutes, and received £2.50 compensation. The experiment was developed with lab.js (Henninger, Shevchenko, Mertens, Kieslich, & Hilbig, 2021), hosted with JATOS (Lange, Kühn, & Filevich, 2015), and the data analysed with jamovi (The jamovi project, 2022).

Bot Accuracy (%)	60	70	80	90	100
Compliance (%)	69	84	85	90	84

Table 1: The proportion of time participants 'complied' with the bot's recommendations, as a function of bot accuracy

Results and Discussion

Results are reported in table 1. As expected, participants made decisions consistent with the robot's recommendations more often when the robot was more accurate. The exception was the initial 100% accuracy block where participants trusted 84% of the time. We attribute this lower trust to participants getting used to the task, and exploring different consequences for their actions.

Participants in the "don't know" condition very rarely made use of that option, apparently preferring to commit one way or the other. 3 of the 4 participants never used the "don't know" option, and the 1 participant who did make use of it, used it less than 10% of the time.

The naive interpretation, that participants were genuinely confident in all the recommendations received seems unlikely, given that, especially towards the beginning of each block, participants had no way of knowing how accurate each robot was. It is particularly surprising that the "don't know" option was not used when the robot was quite unreliable, and was correct only 60% of the time. At that level of accuracy, 60%, the probability that the robot's recommendation being correct is approaching that of a coin-toss; that recommendation therefore has little informative value, and the participant should have very little reason to prefer one outcome over another.

A possible explanation is that the task invites risky decision making. Participants may have found the task boring and were motivated to make it more interesting. Rather than behaving to maximise their fictitious financial reward, a risky "beat the odds" strategy could have made the task more interesting. Participants may have elected to improve their emotional state, rather than optimise their fictional financial reward.

Experiment 2

One limitation of the approach of earlier studies such as Yu et al., and our Experiment 1 replication, lies in the categorical nature of the response. Participants are forced into a choice of trusting vs not trusting. If trust is a matter of degree, the internal continuum cannot be captured by a binary response regime that is common in investigations of human-bot trust.

Another limitation of these earlier studies is that participants typically have no information that would allow them to make that decision without the recommendation. For example, when participants decide whether a light-bulb is broken or not, the only information they have is from the bot. In practice, human-bot teaming involves the human integrating their own judgement with that of the bot, and mediating tension when these judgements diverge.

To enable a continuous measure of trust, and to allow tension between the participant's judgement and that of the recommendation, we adapted a task from the toolbox of social psychologists – the judge advisory system (JAS).

The judge advisor system has been used to explore the way that people integrate advice from one or more additional advisors. It typically involves asking participants to make some sort judgement of a continuous quantity, such as estimating the year when the Suez Canal first opened (Yaniv, 2004), estimating the price of backpack models based on information about their features (Sniezek, Schrah, & Dalal, 2004), estimating the mean annual salary of graduates from different business schools (Soll & Larrick, 2009), or estimating a person's weight from a photograph (Gino & Moore, 2007).

Following an initial judgement, participants are exposed to the judgements of others, before updating (or not) their initial judgement. The extent to which participants revise their estimate toward the advisor's estimate provides a measure of "advice taking". The advisor is analogous to the bot in Experiment 1; advice corresponds to the recommendation, and advice taking corresponds to trust.

The judge advisor system has been used to explore a range of topics in social psychology, investigating the influence of self confidence (Gino, Brooks, & Schweitzer, 2012), confidence of advisor, expertise of advisor (Sniezek & Van Swol, 2001), plausibility of the advice, and a range of other factors on the tendency to make use of advice. An attractive part of the judge advisor system for our present purposes is that it captures a continuous measure of trust. If the participant's second judgement remains unchanged from their first, this suggests a lower level of trust in the advisor, whereas if the participant's subsequent response matches the advisor's advice, this suggests a higher level of trust (Van Swol & Sniezek, 2005). Crucially, a continuum of values fall in between.

Historically, JAS studies have computed the "weight of advice'; as an index of trust behaviour (Harvey & Fischer, 1997; Yaniv, 2004). The weight of advice is calculated for each trial as follows:

$$\frac{B–A}{R–A}$$

where A denotes the participants initial response, R denotes the recommendation, and B denotes the participant's second response (Harvey & Fischer, 1997; Yaniv, 2004). If the participant's second response is the same as their first, $B-A$ becomes zero, and the weight of advice is zero. If the participant's second response is the same as the recommendation, the numerator and the denominator become the same, and weight of advice is one. Similarly, if the participant's second response is half way between their first response and the advice, the 'weight of advice' is 0.5. In this way, the weight of advice captures the degree to which a participant weighs their judgement against the advice given.

It should be noted that the weight of advice is only well defined where the second response falls between the first response and recommendation (producing values between 0 and 1). A convention is to truncate values outside the [0,1] interval to 0 and 1.

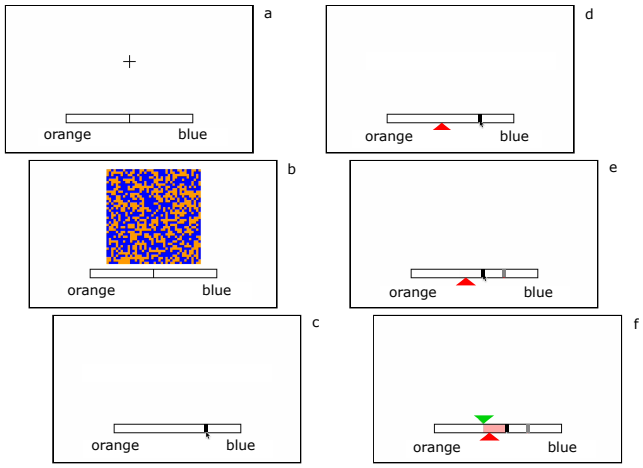

Figure 1: The trial sequence: a. fixation b. stimulus grid is presented c. participant estimates the proportion of colour d. a "recommendation" is provided (the red triangle) e. the participant responds a second time, taking into account the recommendation f. feedback is provided, highlighting the difference between their second response and the true proportion.

More recently, emphasis has shifted from this single index, to a more detailed account. Soll and Larrick (2009) have argued that people use a blend of two strategies. The first is a "choosing" strategy, where participants "take the best" (Gigerenzer & Goldstein, 1996), choosing either their own estimate or the recommendation, and placing their second response on one or the other. This results in a weight of advice of zero or one. The second is an "averaging" strategy, where participants simply take the average of their own and the other judgement (resulting in a weight of advice of 0.5). Soll and Larrick find support for this in a series of experiments where they combine the weight of advice from all trials, across all participants. The resultant histograms from four of their studies each exhibits a trimodal distribution with peaks at 0, 0.5, and 1. In this way the JAS, and examining the distribution of weights of advice can yield deeper insights into people's advice taking behaviour, than simply examining means.

In Experiments 2 and 3 we adapted JAS-like designs to explore the trust of human participants in computer recommendation in a perceptual judgement task. The same paradigm was employed in both Experiments, so we provide a detailed exposition of the method of Experiment 2, and a brief description of the method changes made in Experiment 3.

Method

Design Participants in our study completed a series of simple perceptual judgements, illustrated in Figure 1.

The sequence of events within each trial progress as follows: 1. A fixation cross was presented in the centre of the screen for 500ms (Figure 1a). 2. participants were then pre-

sented with a 20x20 dichromatic array of randomly arranged blue and orange squares, spanning 400x400 pixels (Figure 1b). The display is dynamic such that the position of blue and orange dots constantly changes, at the standard 30 frames per second, but the proportion of colour remained fixed within a trial. 3. The participant then responded by placing a marker on a scale, indicating their judgement of the proportion of colour in the display (Figure 1c). 4. The participant received a recommendation, indicated by a position on the scale (Figure 1d). This recommendation was drawn from a distribution (detailed subsequently), leading to its accuracy varying from trial to trial. 5. Following the recommendation, the participant responded a second time, taking into account (or not) the recommendation (Figure 1e). 6. The true proportion of colour is revealed on the scale, providing feedback as to how close the participant's judgement, and how close the recommendation were to the truth (Figure 1f).

The perceptual decision of determining the proportion of orange and blue pixels would have been trivial if all (or close to all) pixels had been of that colour. After pilot testing we limited the range, such that the left-most end of the scale represented 35% fill of the primary colour (orange or blue, counter-balanced), and the right-most end of the scale represents 65% fill of the primary colour.

Recommendations were generated by combining the true colour proportion with a draw from a uniform distribution of width 10%. In this way, recommendations were always within 5% of the true proportion (and within one sixth of the overall scale).

Although very similar to existing JAS studies, there are a number of differences. Firstly, we make use of a perceptual judgement task. Secondly, the initial judgement, the recommendation, and the subsequent judgement, occur on a trial-by-trial basis. In contrast, JAS studies typically have participants complete all the initial judgements, before providing them with all the recommendations and soliciting revised judgements. Thirdly, we provide feedback throughout the task, allowing participants to develop a sense of their own accuracy, and the accuracy of the recommendation. Finally, we do not frame the recommendations as coming from a human advisor, but rather the task, being rather like a computer game, naturally leads participants to think the recommendations are being generated by the computer (which of course, they are).

The task was completed in two phases, an initial judgement-only phase followed by the recommendation phase. In the judgement-only phase, participants simply performed the perceptual task without receiving a recommendation or needing to respond a second time, giving them the opportunity to become comfortable with the task. In the subsequent recommendation phase participants were presented with the recommendation, and were asked to respond a second time. The complete experiment contained 40 judgement-only trials (20 practice, 20 experimental), followed by 120 recommendation trials (20 practice, 100 experimental).

The experiment was developed with lab.js (Henninger et al., 2021), hosted with JATOS (Lange et al., 2015), and the data analysed with jamovi (The jamovi project, 2022).

Participants 20 participants were recruited via the Prolific online platform. They completed the task in around 20 minutes, and received £2 compensation for their time.

Results and Discussion

Accuracy scores were calculated for each response by subtracting the true proportion from the first and second responses, then taking the absolute values. These values represent the judgement error between participants' responses and the truth. For each participant, their median error values were taken as measures of their individual performance.

Accuracy increased between first responses (mean error 2.31%, SD = .79%), and second responses (mean error 1.87%, SD = .57%). This difference was significant ($t = 4.158, p < .001, BF_{10} = 63$, Bayes factors computed per Morey, Rouder, Jamil, and Morey (2015); Rouder, Speckman, Sun, Morey, and Iverson (2009)). For reference, the recommendations had a mean error of 2.5%. At an individual level, 11 participants achieved significant improvements between their first and second responses at the .05 level (paired wilcox test, one-tailed, adjusted for multiple comparisons). Of these 11 participants, 7 performed better on their first responses than the recommender.

The results suggest that people are able to successfully incorporate the recommendation into their own judgement, in a way that leads to improved performance.

We characterise trial responses as one of five response types (see Figure 2) ; a "stay" response, where the participant responds the second time no different to the first, an "adopt" response, where the participant wholly adopts the recommendation and places their second response on top of it, a "shift" response, where the participant places their second response somewhere between their first response and that of the recommendation, an "overshoot" response, where the participant places their second response on the far side of the recommendation from their first response (corresponding to a weight of advice greater than 1), and a "distrust" response, where the participant places their second response on the far side of their first response from the recommendation (corresponding to a weight of advice less than zero).

Due to the continuous nature of the response scale, it is very unlikely that the second response will fall either exactly on the first response, or exactly on the recommendation. We classified responses as falling within 0.5% of the first response or the recommendation as "stay" and "adopt" respectively. The breakdown of response types for experiment 2 is depicted in table 2.

Four participants made extensive use of "stay" responses, making almost no use of the recommendation. In the majority of responses, they placed their second response almost exactly on their first. This is consistent with having much greater trust in their own ability, and indeed in 3 out of the

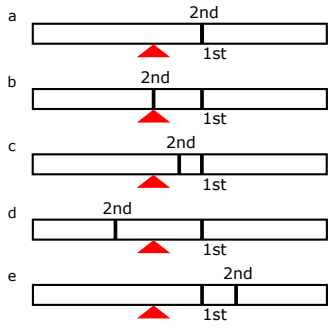

Figure 2: Different response types: a. "stay", first and second responses are the same ($WoA = 0$) b. "adopt", second response is placed on the recommendation ($WoA = 1$) c. "shift", second response falls between the first response and the recommendation ($0 < WoA < 1$) d. "overshoot", the second response goes past the recommendation ($WoA > 1$) e. "distrust", the second response retreats from the recommendation ($WoA < 1$)

	Stay	Shift	Adopt	Overshoot	Distrust
Exp. 2	44.3	36.6	5.6	6.2	7.5
Exp. 3	49.6	29.6	7.1	6.2	7.5

Table 2: The breakdown of different response types (percentages) for experiments 2 and 3

4 cases, participants were outperforming the recommender. One could appreciate them seeing that as they already performed better than the recommender, there was little to gain by paying attention to it. However, it turns out that this is not the case: averaging is almost always a superior strategy to choosing the better estimate (see Soll and Larrick (2009) for an extensive discussion). Of the 11 participants who's accuracy improved between their second and first responses, all made extensive use of "shift" responses, which suggests an averaging strategy.

We aggregated weight of advice of all trials from all participants to produce the histogram in Figure 3. Unlike Soll and Larrick (2009), no peak was found around the value of 0.5, rather the weight of advice for "shift" responses are distributed far more uniform, suggesting participants are not relying on a simple averaging strategy but something with variable weights.

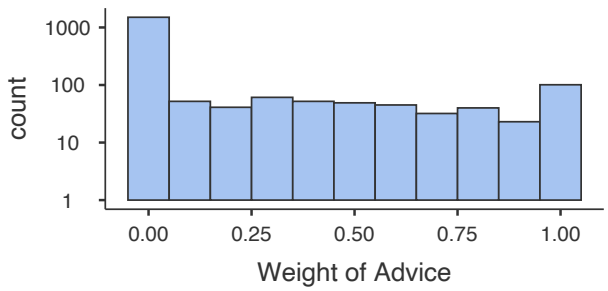

Figure 3: Distribution of Weight of Advice for experiment 2, aggregated across all participants

Experiment 3

Experiment 3 is a variation on Experiment 2, where participants are provided with a recommendation on only 50% of trials. For the trials that do not come with recommendations, the participant is simply asked to respond a second time. If a similar improvement is seen between first and second responses in both trials with recommendations and those without, this suggests the improvement comes simply from making two responses, rather than from the recommendation. Alternatively, if the second responses from trials without a recommendation are not improved, then this suggests that the recommendation is driving the improved performance in trials with recommendations.

Methods

The stimuli and procedure were similar to Experiment 2 with one exception: only half of the trials came with a recommendation whereas the remaining half did not and simply prompted the participant to respond a second time. 20 fresh participants who had not completed earlier experiments, were recruited on prolific and compensated for their time.

Results and Discussion

As before, accuracy scores for each participant were computed. Participants achieved lower judgement errors in their second response (1.92%, SD = .68%) than in their initial response (2.13%, SD = .77%), however this was not significant ($t = 1.926, p = .059, BF_{10} = 1.08$). At the level of the individual, only 3 out of 20 participants demonstrated a significant improvement in performance with the recommender than without ($p < .05$, adjusted for multiple comparisons).

A breakdown of response types for experiment 3 is depicted in Table 2. As can be seen, there was an increase in "stay" responses, and a decrease in "shift" responses compared to Experiment 2. A test of independence revealed a significant difference between experiments 2 and 3 ($\chi^2_4 = 16.16$, p = .003).

The effect of the "no-recommendation" trials, resulting in a reduction in post-recommendation performance was unexpected. Experiment 3 had fewer recommendation trials than Experiment 2 (50 vs 100). The differences between the experiments could therefore be due to sampling error. However, even when considering only the group means, there is a decline in the improvement between experiments regardless.

A possible explanation for the difference is that part of the post-recommendation improvement seen in Experiment 2 actually reflects an improvement in the quality of the second response, over the first. It may be that the participant provides a rushed, low quality response to the first prompt, knowing that they have the opportunity to improve it in the second. The

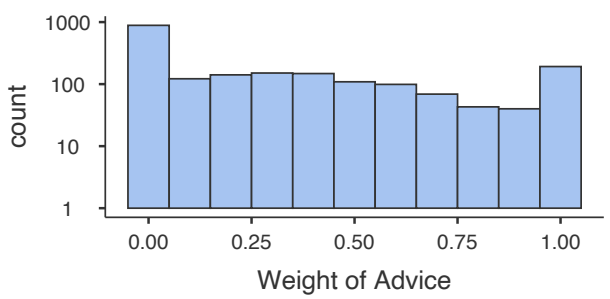

Figure 4: Distribution of Weight of Advice for experiment 3, aggregated across all participants.

second response improves on the first not simply because of the recommendation, but also (or perhaps entirely) because the participant's second response is more careful.

Given that the feedback provided at the end of the trial only highlights the difference between their second response, and the true proportion, it is possible participants perceive this as the important quantity to optimise, and see the first response as irrelevant and not worth putting effort into. Consistent with this account, of the three participants who demonstrated an improvement in performance, two exhibited similar improvements in trials that did *not* involve recommendations.

For the remaining 17 participants who did not exhibit an improvement in performance between responses, we propose that the absence of recommendations on some trials gave the participants less 'cover' to improve their second responses. When participants know they are going to receive a recommendation, they can provide a low quality first response, knowing that they can respond more carefully subsequently, and any improvement can be attributed to the recommendation. In contrast, participants who do not know whether they are going to receive a recommendation need to make a high quality first response, as any improvement can only be attributed to them not applying themselves to the first response.

Consistent with this account accuracy on the first response in Experiment 3 was higher than in Experiment 2 (although not significantly), and participants shifted their second response less often than in Experiment 2.

It may be feared that other judge-advisor system studies experience the same flaw as our Experiment 2, however the protocols in most JAS studies do not inform the participant that there will be an opportunity to revise their first estimate until after they have made it. Unaware that they may respond a second time, participants would not realise they could perform a lazy initial response they can correct later.

The mystery remains, however, why participants in this experiment were unwilling to take advice compared to other studies. One difference between our study and existing JAS studies is that it provides an opportunity for participants to develop a sense of both their own expertise, and the expertise of the recommender. Most JAS studies require the participant to make judgements in areas in which they have little sense of their own expertise. For example, Yaniv (2004) asked participants to estimate the dates of significant historical events, but excluded participants who had majored in history.

Other studies have provided participants the opportunity to develop a sense of their own expertise. For example, Harvey and Fischer (1997) presented a cue-learning task, and included a training stage which provided feedback on their accuracy. Similarly, Gino and Moore (2007) tasked participants with estimating the weight of people from photographs, and provided a training stage with feedback. Although in both these cases participants had opportunity to learn about their own expertise, they had no opportunity to develop a sense of the advisor's expertise to weigh against their own. In the face of such uncertainty, participants may be willing to trust recommendations.

In contrast, the present experiment provided participants time to develop a sense of both their own expertise, and the expertise of the recommender. Informed participants may not be likely to trust recommendations, unless they are substantially better than their own.

Conclusions

We extended two experimental frameworks, the trust dynamics framework of Yu et al. (2017), and the Judge Advisor System. Together, our experiments highlight some of the challenges of evaluating trust in human-bot collaboration. Experiment 1 highlighted the challenges of achieving participant engagement. It seems unlikely that real-world scenarios with stakes high enough to warrant combined human-bot teams, would result in such risky behaviour from human participants. Future research will need to ensure the human is suitably motivated in order to be confident that findings generalise to real world scenarios. Experiment 2 provided some promising results of human-bot collaboration, however Experiment 3 suggested that much of the human-bot teamwork illusory, and participants were largely working on their own.

The chief finding is that people exhibit much lower levels of trust in the provided advice, or recommendations, than has been observed in earlier Judge Advisor System studies. The reason for this is not entirely clear, but it may be that existing JAS studies consider scenarios where people do not have enough information to meaningfully weigh their own expertise against the advisor or recommender. People may be more inclined to "hedge their bets" in such situations, whereas perhaps they may have been content with their own judgement had they realised the recommender or advisor was not substantially better than them.

Armed with a continuous measure of trust (albeit one that requires further testing), future studies can yield rich data supporting development of formal models of human-AI trust.

Acknowledgements

This research was supported by AUSMURI and ARC-DP grants to SB and AE (AUSMURIIV000001 and DP210100313), and by an Australian Government RTP Scholarship to JL.

References

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 7, pp. 2–11).

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, *103*(4), 650.

Gino, F., Brooks, A. W., & Schweitzer, M. E. (2012). Anxiety, advice, and the ability to discern: feeling anxious motivates individuals to seek and use advice. *Journal of personality and social psychology*, *102*(3), 497.

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, *20*(1), 21–35.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, *70*(2), 117–133.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2021). lab. js: A free, open, online study builder. *Behavior Research Methods*, 1–18.

Lange, K., Kühn, S., & Filevich, E. (2015). " just another tool for online studies"(jatos): An easy solution for setup and management of web servers supporting online studies. *PloS one*, *10*(6), e0130834.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*(1), 153–184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50–80.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'bayesfactor'. *URLh http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)*.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, *27*(5-6), 527–539.

Razin, Y. S., & Feigh, K. M. (2023). Converging measures and an emergent model: A meta-analysis of human-automation trust questionnaires. *arXiv preprint arXiv:2303.13799*.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*, 225–237.

Sniezek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, *17*(3), 173–190.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes*, *84*(2), 288–307.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 780.

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). Ai-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, 1–18.

The jamovi project. (2022). *jamovi.* Retrieved from `https://www.jamovi.org`

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British journal of social psychology*, *44*(3), 443–461.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes*, *93*(1), 1–13.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 307–317).