AFRL-AFOSR-UK-TR-2023-0035

Decision Confidence in Human-Machine Teaming

Nick Yeung
THE UNIVERSITY OF OXFORD
UNIVERSITY OFFICES
OXFORD, ,
GB

12/22/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20221222	Final	START DATE 20180501	END DATE 20220430

4. TITLE AND SUBTITLE
Decision Confidence in Human-Machine Teaming

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-18-1-0207	
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Nick Yeung

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
THE UNIVERSITY OF OXFORD
UNIVERSITY OFFICES
OXFORD
GB

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2023-0035

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project aimed to characterise and increase trust in human-machine teaming by leveraging insights from research on trust and influence in human social and group decision making. We conducted two complementary streams of research to investigate (1) the causes and consequences of algorithm aversion, whereby human operators systematically downweight advice from artificial systems after seeing them err; and (2) the features of algorithmic design that promote trust in advice from artificial systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	c. THIS PAGE		
U	U	U	SAR	18

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
NANDINI IYER	314-235-6161

Final (Y4) Report for EOARD Grant FA9550-18-1-0207

Decision confidence in human-machine teaming

October 2022

Name of Principal Investigators (PI and Co-PIs): Prof Nick Yeung
- e-mail address : nicholas.yeung@psy.ox.ac.uk
- Institution : Department of Experimental Psychology, University of Oxford
- Mailing Address : Anna Watts Building, Oxford, OX2 6GG, UK
- Phone : +44 (0)1865 271389

Period of Performance: 5/1/2018 – 4/30/2022

Team: Nick Yeung, Lead investigator; Sriraj Aiyer, Research Assistant; Lt Col Aaron Celaya, doctoral student.

Abstract: This project aimed to characterise and increase trust in human-machine teaming by leveraging insights from research on trust and influence in human social and group decision making. We conducted two complementary streams of research to investigate (1) the causes and consequences of algorithm aversion, whereby human operators systematically downweight advice from artificial systems after seeing them err; and (2) the features of algorithmic design that promote trust in advice from artificial systems.

Accomplishments

Research objectives

Human-machine teaming is becoming increasingly important in today's world. However, in contexts as diverse as political forecasting, school admissions and even day-to-day GPS navigation, human operators have been shown to make sub-optimal use of information provided by intelligent machines such as expert systems and algorithm-based predictions: In some cases, human decision makers systematically ignore or fail to act on good-quality information provided by artificial systems ("algorithm aversion"); in other cases, they show over-compliance or complacency in relying on poorly performing artificial systems. The causes and prevalence of these diverging patterns of algorithm aversion vs. over-trust remain poorly understood. This project investigated these issues in two complementary streams of research. First, we investigated the causes and consequences of algorithm aversion, whereby human operators systematically downweight advice from artificial systems after seeing them err. Second, we investigated the features of algorithmic design that promote trust in advice from artificial systems.

The core motivating assumption of our project was that people's trust in algorithms follows the principles of trust that underpin human social interaction. As such, insensitivity to mechanisms of interpersonal trust will undermine the effectiveness of human-machine teaming. For example, communication of confidence critically underpins human trust: Confidently expressed opinions have more influence, but trust is lost if this confidence proves to be unfounded; more confident decision makers are less receptive to advice and down-weight dissenting opinions; and teams can make optimal decisions that outperform the best individual decision makers only if team members communicate confidence effectively.

Human-machine teaming may remain critically limited if such factors are ignored or lacking in system design.

Technical approach

Our experiments used well-characterized perceptual decision making tasks, in which the evidence used for decisions can be carefully controlled and for which normative models of decision making have been developed (Boldt & Yeung, 2015; Yeung & Summerfield, 2012). We have used these tasks successfully in previous research to investigate human trust and influence (Carlebach & Yeung, 2022; Pescetelli, Hauperich & Yeung, 2020; Pescetelli & Yeung, 2021).

A schematic of the task structure used throughout our research is shown in Figure 1. In the basic task, participants judge which of two briefly-presented boxes contains more dots, with task difficulty precisely controlled via the relative numbers of dots presented in each. After registering an initial response, indicating which box contains more dots and rating their confidence in this judgment, the participant is offered advice, for example choosing between receiving advice from a human or computer advisor. Following the advice, they have the opportunity to revise their initial answer on the same sliding confidence scale. At the end of the trial, they may be given objective feedback. Participants perform repeated trials of this task, and therefore have the opportunity to learn about the quality of advice they receive from different advisors, and to change their selection and use of this advice accordingly. We assess trust in advice and advisors in three ways: the degree to which participants update their decisions and confidence as a function of the advice provided (advisor influence); post-task ratings of advisor accuracy, reliability and trustworthiness (advisor ratings); and choice of which advisor to receive advice from (advisor choice).

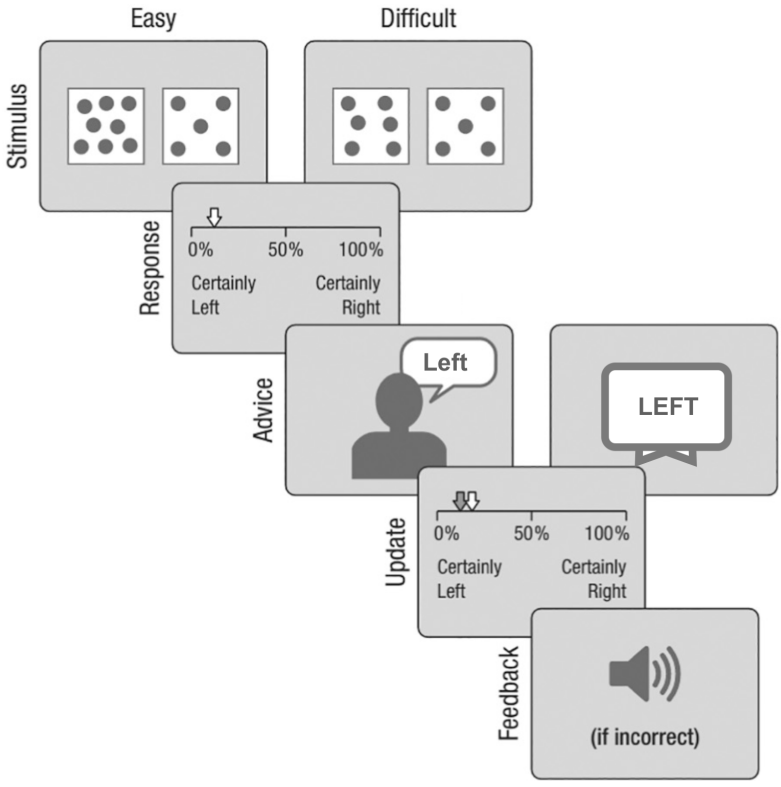

Figure 1. Schematic of experimental paradigm. Participants performed perceptual judgments with the opportunity to choose and use advice. Note that the actual dot task stimuli used in the experiment were more detailed than is depicted here: They contained 200+d and 200-d

dots, where d was an experimental parameter that varied to control the difficulty of the perceptual decision.

Human advice was derived from a database of stimuli and real decisions from an initial round of data collection with the perceptual decision task. Computer-generated advice was derived from a sampling algorithm that based its perceptual decision on a subset of available data (i.e., a sub-region within the box of dots). The size of the sampling subset can be varied systematically to produce a specific desired level of accuracy for algorithmic advice, e.g., that can be matched with human advice. Across experiments we investigated task and individual-difference factors that affect participants' trust in advisors.

Major activities

We conducted two major lines of research. The first focused on factors influencing relative trust in human vs. algorithmic advice. Our initial hypothesis was that we would observe algorithm aversion in our decision making task, such that participants would systematically prefer human over algorithmic advice even when the advisors were matched in their objective accuracy. However, in our experiments we did not find evidence of algorithm aversion, and indeed found a slight overall preference for algorithmic advice in our task. Instead, the most striking feature of our data was wide individual variability in preference for human vs. algorithmic advice. Subsequent experiments focused on characterising these individual differences. This work led to the development of new survey tools to assess individuals' trait-level trust in algorithms. The major activities in this line of research comprised (a) running experiments to collect behavioural data in our task paradigm, initially from in-person testing in our lab at Oxford, and subsequently in online samples; and (b) psychometric data analysis for survey tool development.

Our second line of research investigated factors promoting trust in algorithmic advice. This comprised two series of experiments. The first series explored preferences for different kinds of algorithmic advice. A key intuition about algorithmic-based advice is that it should provide answers with certainty. However, although human decision makers are typically more influenced when advice is provided confidently (Yaniv, 2004), they can show greater trust when advice acknowledges the uncertainty that is inherent in many complex decision making contexts (Gaertig & Simmons, 2018). Our first question was therefore whether human decision makers would prefer algorithmic advice provided with certainty (simple binary decision) or with accompanying information about its likely reliability (confidence). In the second series, we studied the impact of timing of advice: whether advice is more effective when provided before vs. after participants had the opportunity to evaluate evidence themselves. Although several lines of reasoning would suggest that people should prefer advice beforehand, as a shortcut to making the decision themselves and because of cognitive biases such as confirmation bias, it is also possible that later advice is more effective as it can be integrated with the participants' own already-formed judgments. The major activities in this line of research comprised two series of experiments and accompanying data analysis.

Specific objectives

1a) Characterise algorithm aversion in our task paradigm.
1b) Investigate determinants of preferences for human vs. algorithmic advice.
2) Identify factors promoting trust in algorithmic advice.

Significant results and key outcomes

1a) Characterise algorithm aversion in our task paradigm

In contrast with our initial expectations, we found no systematic mistrust of computer advice (algorithm aversion) in our decision task, in contrast to some previously published research (Dietvorst et al., 2015) but consistent with research that has emerged in parallel with our own (e.g., Logg et al., 2019). Instead, we found wide variation in participants' reliance on human vs. computer advice, with some participants exhibiting strong preferences for human advice and others strong preferences for algorithmic advice. In part, this variation reflected participants' early experience with the advisors, even when the overall quality of advice from human and computer advisors was perfectly matched (Figure 2). However, in subsequent experiments, we observed large variations in human vs. computer advice preferences even when controlling for this kind of early experience (Figure 3).

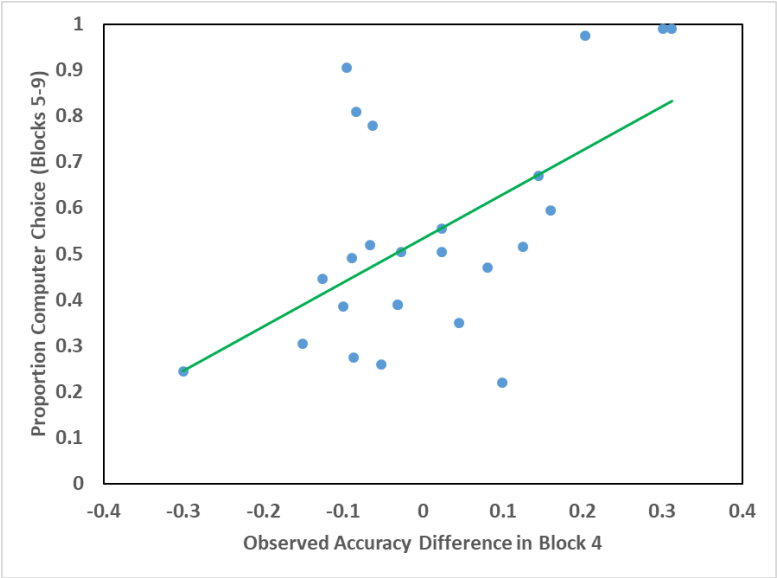

Figure 2. Early Experience and Advisor Selection. Each datapoint represents a single participant. Values along the x-axis indicate observed accuracy difference between two advisors in the first 60 trials of the task with advisors (Block 4, with blocks 1-3 of the task being practice blocks): Positive values along the x-axis indicate the computer advisor performed more accurately, as observed by the participant, whereas negative values along the x-axis indicate the human advisor was more accurate. Values on y-axis indicate the proportion of choice of computer advice (vs. human advice) in the 5 subsequent blocks.

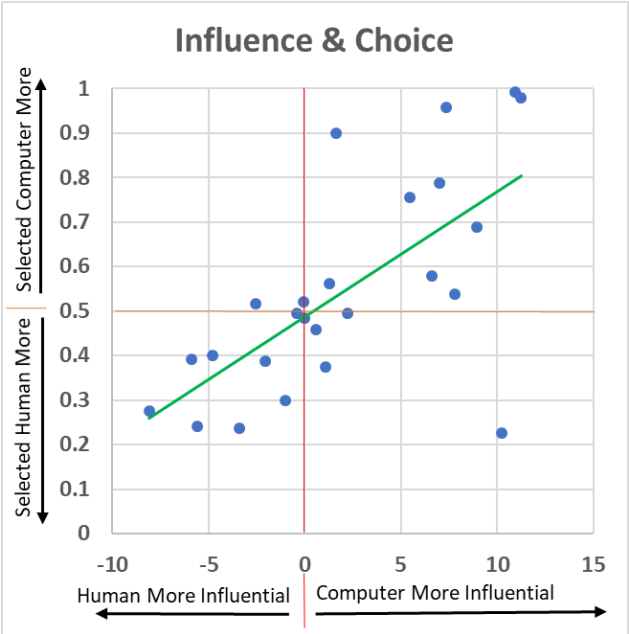

Figure 3. Individual differences in preferences for human vs. algorithmic advice. Each datapoint represents a single participant. Values along the x-axis indicate observed differences in influence of human vs. computer advice: Influence score for each advisor is calculated as the sum of two measures—the degree to which participants increase confidence when the advisor agrees with their initial judgment vs. decrease their confidence when the advisor disagrees. The y-axis plots the difference in proportion of trials for which participants chose algorithmic vs. human advice. The measures are positively correlated, indicating stable preferences evident in the two measures. Critically, there is wide scatter (i.e., large individual differences) along both axes.

Our subsequent work in this stream focused on capturing these individual differences in trust. A first experiment assessed whether preferences remain stable across three different pairings of human vs. computer advisors in the perceptual decision task described above, and whether preferences in the experimental task correlated with participants' trust in artificial systems as revealed via a simple questionnaire in which participants rated their agreement with statements including: "Driverless cars will never be as safe as cars driven by people" and "When on social media, I primarily read news articles that have been recommended to me by the site rather than by seeing what my friends are reading". Within the experimental task context, participants' behaviour was highly regular such that participants who preferred the human advisor over one computer advisor tended to show a similarly strong preference for a second human advisor over a second computer advisor (Figure 4, left panel). In contrast, advisor preferences in the experimental task were, surprisingly, negatively correlated with their questionnaire responses (Figure 4, right panel). We speculated that this negative correlation is indicative that participants with greater experience of algorithms in everyday life were less willing to trust specific algorithms that were unable to outperform human advisors. However, detailed analysis of the questionnaire results (e.g., via Cronbach's alpha) suggested that the internal validity (consistency) of survey items was too low to draw meaningful conclusions.

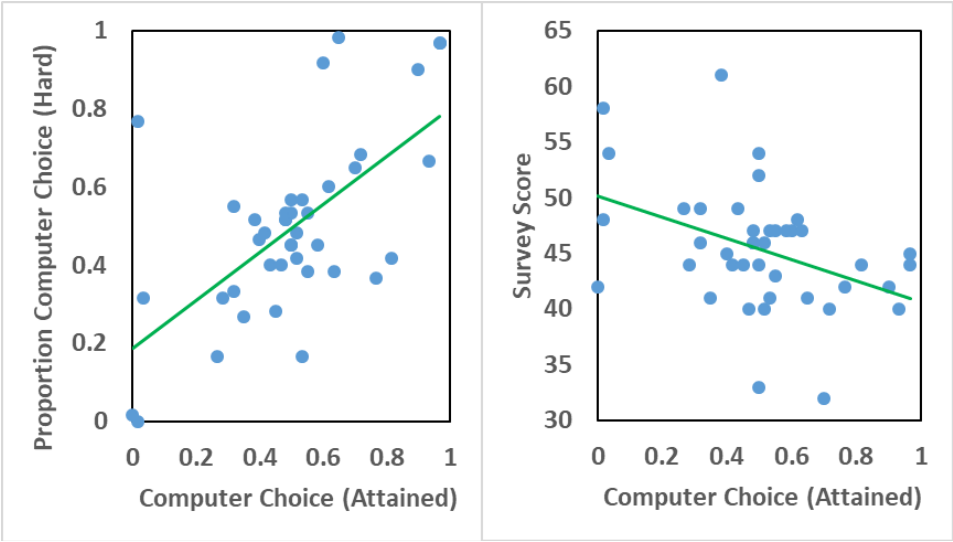

Figure 4. Stability and predictability of individual differences in preference for human vs. computer advice. Left panel: Relationship between proportion computer advice choices in standard ("attained") difficulty blocks of the perceptual decision task (x-axis) vs. computer advice choice in blocks when the perceptual task is more difficult (y-axis). There is a strong positive relationship indicating stable preferences for computer vs. human advice across advisor pairings. Right panel: Relationship between proportion computer advice choices in standard difficulty blocks of the perceptual decision task (x-axis) vs. survey score of computer trust in everyday contexts (y-axis). There is significant negative correlation between these measures, contrary to our original predictions.

1b) Investigate determinants of preferences for human vs. algorithmic advice

Based on the observations above, we devoted effort towards developing new survey tools to assess individual differences in trust in automation and intelligent systems. The need for such tools is well illustrated by a paper from the Australian Government, which noted that "future work on the human side of human–autonomous system interactions should include the development of a psychometrically reliable and valid instrument for measuring attitudes and beliefs about the general propensity to trust automated and autonomous systems." (Davis, 2019, p.28). Existing scales have a number of issues that have hindered their widespread adoption. Surveys on automated and intelligent systems can become outdated due to the rate of technological progress, affecting their long-term utility. For example, a widely-cited scale from Singh et al. (1993) mentions outdated technologies such as VHS and library card catalogues. Many other scales in widespread use (e.g. Davis's technology acceptance scale) measure people's attitudes to a particular piece of technology, rather than their general attitudes to AI and automation.

In a first stage of survey development, we collated survey items to create a large corpus of 412 items, primarily based on previously-reported surveys and questionnaires but supplemented with further items of our own (including from the questionnaire above). Removing duplicates and items irrelevant to a general measure of trust, resulted in a list of 131 items. We split these 131 items into two separate groups according to whether they probed attitudes to intelligent systems in general (84 items) or perceptions of specific instances of intelligent systems (47 items). The former group probed the utility, usage, safety, reliability and perceptions of systems based on past experience without asking about one type of system (e.g., "If I am not sure about a decision, I trust that an intelligent system will provide the best solution."). The latter group contained specific, modern instances of

intelligent systems being used in industries such as medicine, voice assistants and finance (e.g., “Driverless cars, controlled by systems, will never be as safe as cars driven by people”).

Intelligent Systems Survey tool

We administered the 84 items probing general attitudes towards intelligent systems to a sample of 200 online participants. Their data fed into an Exploratory Factor Analysis using Maximum Likelihood (ML) and Principal Axis Factoring (PAF) as extraction methods. For this initial stage, no assumptions were made that existing models of trust (e.g. Mayer et al., 1995 Hoff and Bashir, 2015) would be evident in our results. Without an a priori hypothesis of a factor structure, we used EFA at this stage of our survey development. We developed two different factor models in tandem, each with a different set of questions (Figure 5). One breaks down trust into three constituent parts while the other maintains a more general sense of trust. Of particular interest, for the 3-factor model, we found that the questions clustered in a way that corresponds well with influential theoretical models of trust that emphasise separable contributions of ability and integrity, with an additional cluster relating to people’s familiarity with or understanding of intelligent systems (Figure 6).

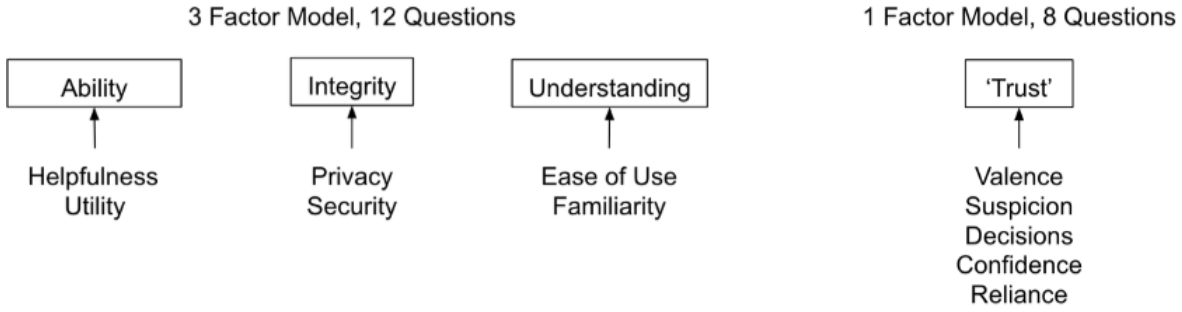

Figure 5. Summary of our two proposed Intelligent Systems Survey tools and the topics covered in their questions

Ability	**I would feel a sense of loss if an intelligent system was unavailable and I could no longer use it.**
	Intelligent systems help me to understand the world around me.
	Intelligent systems help me to plan.
	I spend a lot of my time interacting with intelligent systems on an average day.
Integrity	**I'm concerned that intelligent systems will collect too much personal information from me.**
	I worry that intelligent systems will lead to a surveillance state.
	I'm worried about the general safety of intelligent systems.
	I am suspicious of an intelligent system's action, or outputs.
Understanding	**I have the knowledge necessary to use intelligent systems.**
	I find it easy to learn to use new intelligent systems.
	Intelligent systems are difficult to navigate.
	I find intelligent systems easy to use.

Figure 6. Questions from our three-factor model Intelligent Systems Survey and the topic names used to describe the general theme conveyed by each group of questions.

We assessed the reliability of the revealed factor structure of the two surveys via a round of confirmatory factor analysis (CFA). To this end, each of the two survey tools (12- item/3-factor and 8-item/1-factor) was administered to samples of approximately 300 new online participants. We applied CFA to the resulting data using maximum likelihood estimation with bootstrapping (with 1000 resamples) and multiple model fit measures including Chi-Square, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Standardized Root-Mean Square

Residual (SRMR) and Root-Mean-Square Error of Approximation (RMSEA). The results indicated intermediate to good robustness of the 3-factor survey, with slightly weaker results for the 1-factor survey.

Evaluating the Intelligent Systems Survey tool

Given its good psychometric properties, we next tested the predictive validity of our 3-factor Intelligent Systems Survey tool in our experimental task. In this online experiment, participants completed the Intelligent Systems Survey followed by two experimental tasks (order counterbalanced across participants): the perceptual decision task with advice as above, and a second task taken from an influential paper on algorithm aversion (Dietvorst et al, 2015), based on experimental materials published later by the same authors (Dietvorst et al, 2018). In this task, participants had to predict the performance of prospective business students on a standardized mathematics test, based on set of information about the student, with the option of delegating answering to an algorithm. Overall, our results indicated positive evidence that the survey could be used to predict behavior on our perceptual decision task (Figure 7), with people scoring higher on trust in the Intelligent Systems Survey being more influenced by algorithmic compared to human advice. Further analysis indicated that the different subscales in our survey may be measuring distinct constructs of trust, as only one subscale displayed a significant correlation with relative influence that was driving the overall effect of the survey—interestingly this was the Integrity subscale (Figure 8). Results for Dietvorst et al.'s estimation task were much noisier, with advice use and delegation choices not correlating significantly with each other, with survey scores, or with behavior in our perceptual decision task, and as such appeared to be an unreliable measure (in terms of stable individual differences). We subsequently conducted a replication study focusing on our perceptual decision task alone, and found comparable results with Intelligent Systems Survey scores significantly predicting the relative influence of human vs. algorithmic advice.

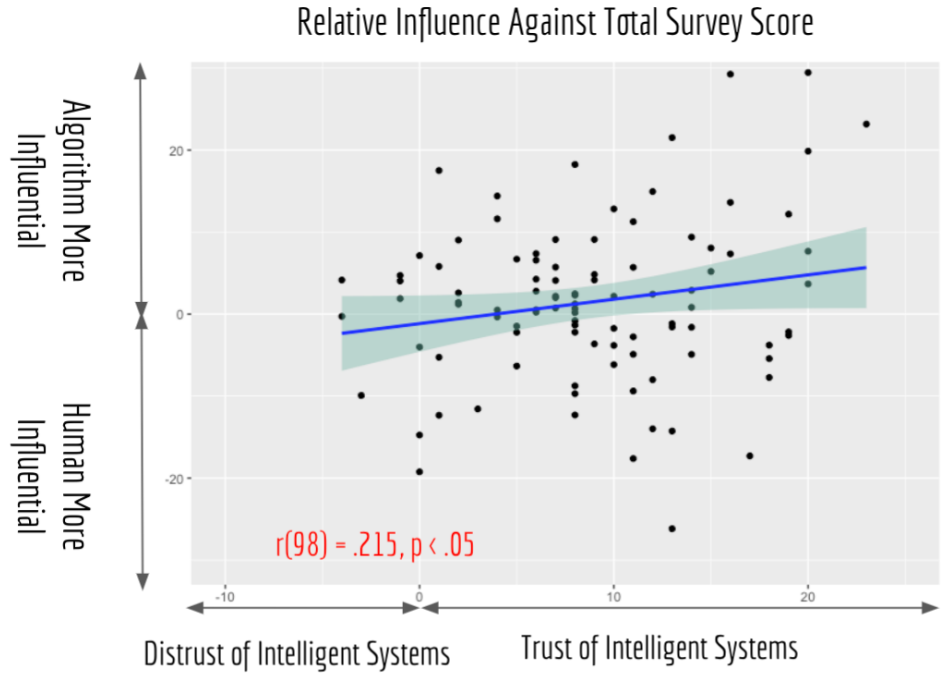

Figure 7. Correlation Between Intelligent Systems Total Survey Score and Behavioral Variable of Influence our perceptual decision task.

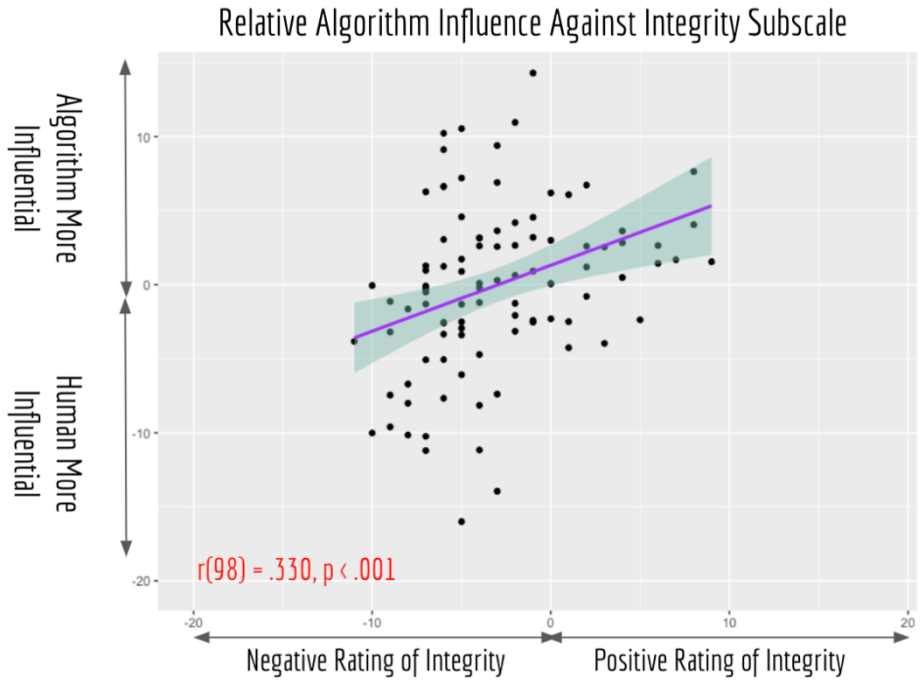

Figure 8. Correlation Between Intelligent Systems Integrity Subscore and Behavioural Variable of Influence from our perceptual decision task. For the other two subscales, the correlation was not significant.

Specific Systems Survey tool

To complement the Intelligent Systems Survey tool, we developed a second survey tool for capturing individual differences in propensity to trust algorithms, via assessing attitudes towards specific instances of technologies. In past work with a similar approach, this has included now outdated technologies such as VHS and card catalogues (Singh et al, 1993). However, there may be value in asking participants for their perceptions of specific kinds of technology, as contemporaneous examples may inform their trust of systems more broadly. We followed the same development procedure as described above for the Intelligent Systems Survey tool. This process resulted in a 3-factor survey with robust psychometric properties, with questions as below.

Transport	Truck driving should be left to humans rather being handled by a system.
	Deliveries of products to customers can only be done by humans.
	I would not use a self driving driverless car as my primary means of transportation.
Decisions	Final decisions about who to hire for a job should be made by qualified people and not by automated hiring systems.
	When sentencing persons convicted of crimes automated systems should be used instead of human judges or juries.
	Automated education systems are more effective than human led classrooms.
Social	Music streaming applications have the ability to make music recommendations for me that I will enjoy.
	When shopping online personalized computer generated ads are very helpful recommendations.
	Social media feeds are useful for suggesting content that I would like

Figure 9. The 9-item Specific Systems Survey tool, with items grouped into three themes of Transport, Decisions and Social applications of intelligent and automated systems.

Capturing individual differences in algorithmic trust

As a capstone to this series of studies, we ran a large sample (N=321), pre-registered (https://osf.io/6nf9j) online experiment in which we collected data from our two surveys, our behavioural task, and two other tasks taken from the literature for which openly shared

paradigms and materials were available. The first of these tasks is an investment task (Fenneman et al 2021), in which the key dependent variable is the difference in the proportion of trials where participants invest in an algorithmic fund manager relative to a human fund manager. The second is a vignette task, where participants are asked to report the extent to which they prefer a human or artificial intelligence to make decisions in a variety of contexts including assigning first-responders to emergency situations and determining the recipients of limited vaccine supplies during flu season (Nussberger, 2021). Here the key dependent variable is Total Risk Score (for relying on AI to make decisions). Of interest was the degree to which we would observe individual differences that were consistent across these measures. Data were collected across two sessions split 2-weeks apart, with both surveys administered separately in each session, enabling us to assess the test-retest reliability of the tools.

Analyses are ongoing as we work towards publication, but initial results are promising albeit somewhat mixed across measures. Although we observed an overall positive correlation between Intelligent System Survey score and relative influence of algorithmic vs. human advice, the effect was not statistically significant ($r(319) = .05$, $p = .41$, Figure 10), in contrast with our initial experiments as described above. However, score on the Intelligent System Survey reliably predicted the relative proportion of choosing algorithmic vs. human advice in our perceptual decision task ($r(319) = .14$, $p = .01$) and Total Risk Score in the vignette task ($r(319) = .12$, $p = .03$), as well as with Relative Investment in the investment task, although the effect was unexpectedly in the opposite direction to that predicted ($r(319)= -.14$, $p = .01$). We ran corresponding analyses for the Specific Systems Survey, finding a marginally significant correlation with Relative Influence of algorithmic vs. human advice in the perceptual decision task ($r(319) = .11$, $p = .06$), a strong correlation with Total Risk Score in the vignette task ($r(319) = .43$, $p < .001$), and a non-significant correlation with Relative Investment ($r(319) = .07$, $p = .21$).

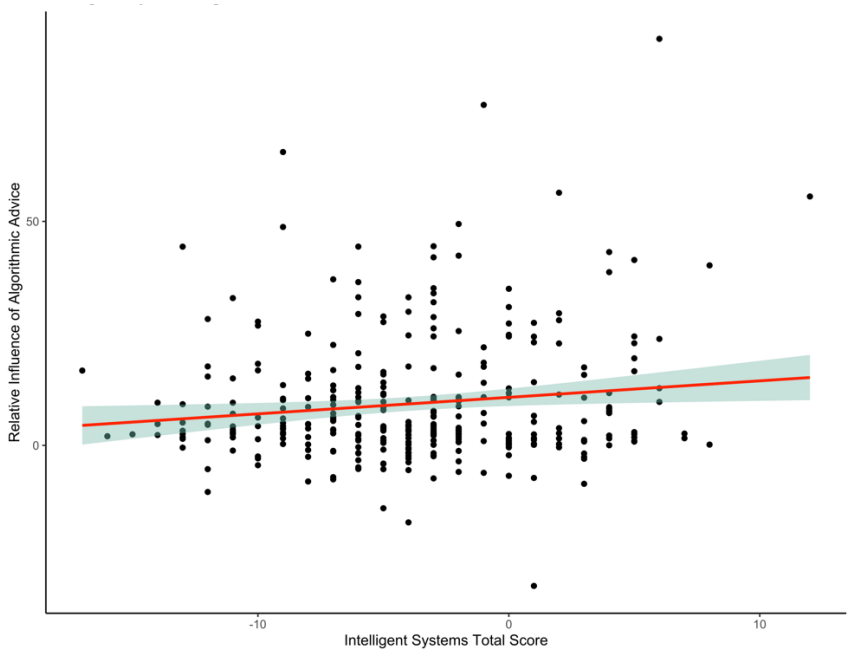

Figure 10. Correlation between the total score on the Intelligent Systems Scale (x-axis) and Relative Influence (y-axis) in the perceptual decision task.

We investigated whether there was common variance between the analogues of trust across tasks using Structural Equation Modelling (SEM). Through this process, we define a latent variable using these observed behavioural variables and then conducted a regression on the scale scores as predictors of this composite (across-task) measure of trust in algorithms. An SEM including Relative Trust in the perceptual decision task, together with Relative Investment and Total Risk Score, yielded a marginally significant correlation with score in the Intelligent System Survey tool ($z = 1.75$, $p = .08$). Using algorithm choice rather than relative trust in the SEM yielded a significant correlation ($z = 2.27$, $p = .02$). Taken together with the individual correlation results, the effect sizes are small but indicate a latent trait-level factor of trust in algorithms that significantly predicts behaviour across different task contexts.

Our analyses of test-retest reliability of our survey tools produced positive results. For the Intelligent Systems survey, we found a correlation over time of .76 on individual item scores and a spearman correlation on the total scores of .62 ($p < .001$). For the Specific Systems survey, we found a correlation over time of .80 on individual item scores and a spearman correlation on the total scores of .69 ($p < .001$). These results indicate significant test-retest reliability for both surveys.

In a final analysis, we evaluated overall patterns of trust across each of the tasks. As observed in previous experiments, we found an overall strong preference for algorithmic advice on the perceptual task as more participants chose to receive algorithmic advice on more trials compared to human advice (Algorithm Choice Mean = 0.64, SD = 0.21). There was a weak preference for the algorithmic fund manager on the investment task (Relative Investment Proportion Mean = 0.02, SD = 0.15), whilst there was an overall preference for human decision makers on the vignette task (Total Risk Score Mean = -1.10, SD = 11.33). This indicates very different assignments of algorithmic trust depending on the task context.

2) Identify factors promoting trust in algorithmic advice

This line of work investigated factors promoting trust in algorithmic advice, focusing on whether people express preferences for (1) simple binary vs. graded advice; and (2) advice received prior to vs. after their own decisions.

Simple vs. graded advice

Based on the intuition that algorithmic-based advice should provide answers with certainty, we contrasted preference of algorithms that provide simple (binary) advice vs. an algorithm that provides accompanying information about the likely reliability of its recommendations (confidence). A first experiment (N = 40) compared human participants' preferences for different types of algorithmic advice in our perceptual judgment task. Both advisors were algorithms (in contrast to the experiments above that compared human vs. computer advice), with the two algorithms differing crucially in providing either simple binary advice (i.e., advice that left or right box is the correct answer) or advice with graded confidence (i.e., advice that left/right box is correct with an associated % certainty in this recommendation). The results indicated that participants on average preferred the advisor providing graded confidence judgments ($t(39) = -3.97$, $p < .001$, Figure 11), which is normatively justified given that this advisor conveys more information. This finding suggests that human decision makers do not show aversion to algorithms that express advice with uncertainty, consistent with what they show for human advice (Gaertig & Simmons, 2018).

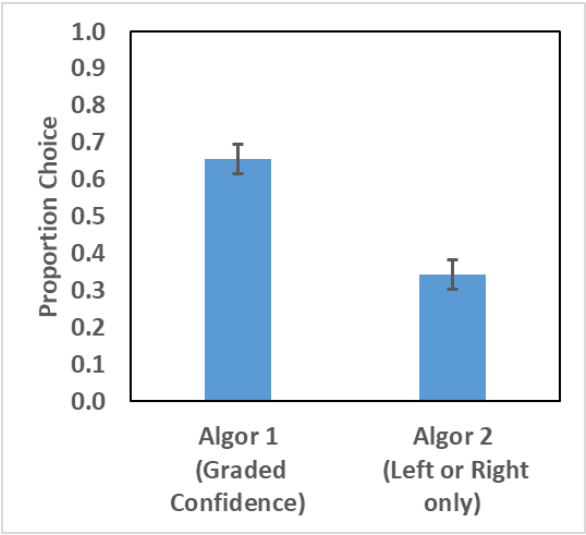

Figure 11. Preferences for different types of algorithmic advice. Proportion choice of two algorithms in Experiment 2a, showing an overall preference for an algorithm that provides information about the likely reliability of its advice (graded confidence) over simple binary advice (left or right box contains more dots).

We conducted two follow-up experiments to assess the stability of participants' advisor preference. The experiments varied cognitive load and speed/accuracy pressure in the task, respectively. The rationale for both studies was that both manipulations might be expected to increase participants' preference for simple, binary advice. Thus, in each experiment, after two initial experimental blocks that followed the design of previous studies, participants performed the task for 2 blocks across which we varied either cognitive load (concurrent working memory task with low or high amounts of information to retain) or speed-accuracy pressure (emphasis on fast vs. accurate responding). In both experiments, we observed an overall preference for graded advice that was stable across the manipulations, suggesting that the preference for the richer source of advice information is a consistent one. Across all three experiments combined, there was a strongly significant preference for the algorithmic advisor that gave an reliability estimate for the quality of advice provided.

Advice order

A series of three online behavioural experiments investigated the relative impact of advice that is received prior to vs. after participants themselves had the opportunity to view evidence relevant to a decision. The experiments used the same perceptual decision task as above. The critical independent variable was whether advice preceded or followed presentation of the dot stimuli on each trial, which varied across blocks (Figure 12). We additionally varied the reliability of the advice—across blocks for each participant, the advice varied between 50%, 70% and 90% accuracy—and whether trial-by-trial objective feedback was provided (varied between participants). This design enabled us to evaluate the impact of advice order as benchmarked against the effect of objective advice quality.

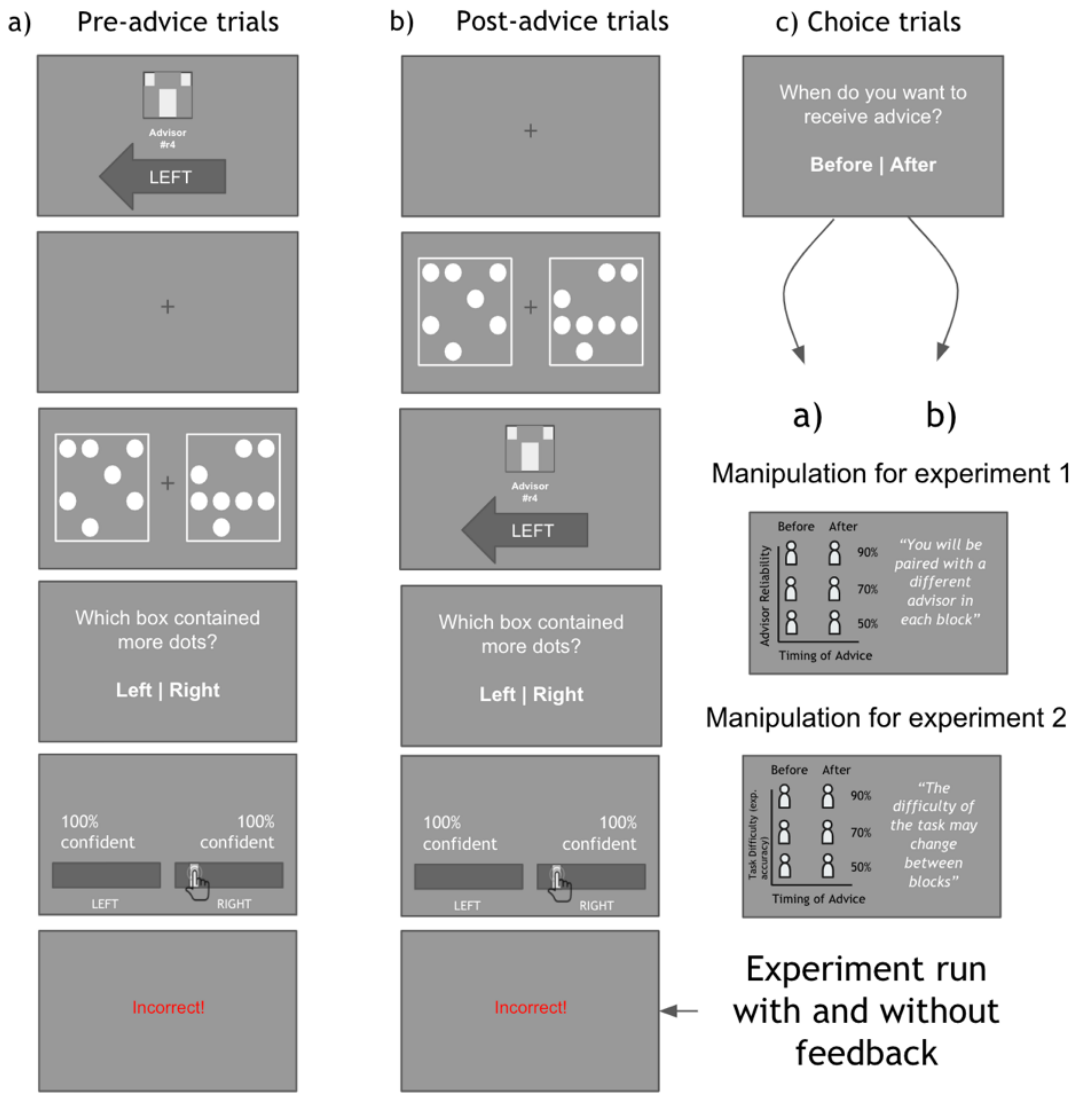

Figure 12. Paradigm for advice order experiments. Panels (a) and (b) depict the sequence of events within a single trial. Panel (a) depicts a pre-advice trial, which begins with a virtual advisor (pictorial avatar) giving advice on the correct answer (here: LEFT), followed by a brief fixation screen and then the dot stimulus, to which participants respond by indicating their judgment (left/right box has more dots) and associated confidence on a visual analogue scale (where clicks on more extreme values indicate higher confidence). Some participants were given trial-by-trial feedback on each judgment. Panel (b) shows the corresponding sequence for a post-advice trial. In some experiments, participants were given a choice of advisors on a subset of trials – panel (c). The inset panel on the lower right summarises the experimental manipulations.

The results of the three experiments indicated a consistent effect whereby post-evidence advice had a greater impact on participants' decisions. For example, as shown in Figure 13, participants were more likely to agree with post- than pre-evidence advice and, as shown in Figure 14, were more influenced by post-advice as shown in differences in confidence between decisions that agreed vs. disagreed with advice received. Figures 13 and 14 also show, as expected, that participants were sensitive to objective advice reliability.

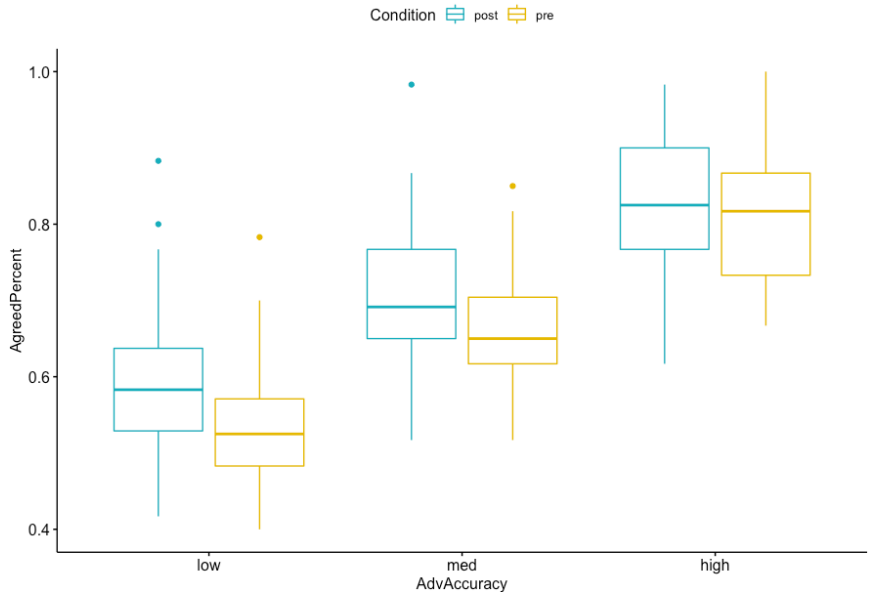

Figure 13. Rate of agreement with advice as a function of advice order and advisor accuracy (reliability). This analysis is taken from the dataset where participants saw feedback. Similar (but numerically larger) effects were seen for participants who did not receive trial-by-trial feedback.

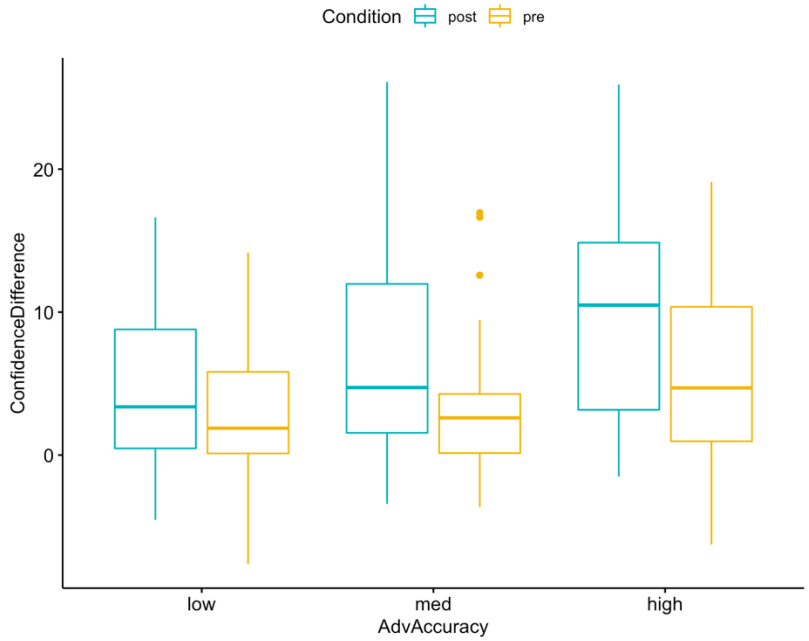

Figure 14. Influence of advice, with the y-axis values plotting the difference in participants' confidence when they agree vs. disagree with the advice given. This analysis is taken from the dataset where participants saw feedback. Similar (but numerically larger) effects were seen for participants who did not receive trial-by-trial feedback.

The second experiment of the series replicated this differential impact of advice as a function of order, and showed that it held regardless of variations in task difficulty. In addition, in this experiment, participants had the opportunity to choose on some trials whether to receive advice before vs. after they viewed the stimulus, enabling us to assess explicit preferences as well as differences in the influence of advice. The results showed further that participants

were not only influenced more by post-advice, but also exhibited an explicit preference for this type of advice when given the choice (Figure 15). The final experiment again replicated the preference for post-advice in both the standard version of our task (a binary judgement of whether there were more dots in a left or right-hand box) with those observed in a perceptual estimation task on the same kind of visual stimuli. This new task required participants to estimate a quantity on a continuous scale: how many more dots they saw in one grid compared to the other. The rationale for this additional condition was that "anchoring" effects of early advice (whereby external information has a disproportionate impact on a person's judgments, acting as an anchor that they struggle to adjust away from) might be much stronger in estimation tasks than in categorical judgments. However, once again we found a larger impact of advice received after participants had themselves viewed the stimulus.

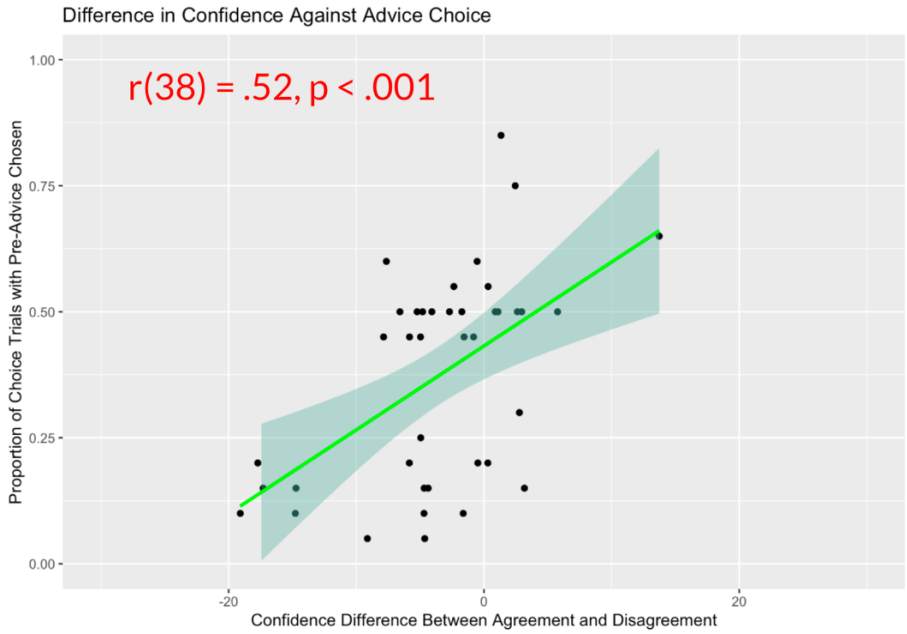

Figure 15. Advice-order choices and influence. In two blocks of trials, participants chose every 6 trials whether they wanted to receive advice before or after they saw the stimulus. This graph plots the proportion of trials where pre-stimulus advice was chosen by the participant (y-axis) against the relative influence of advice for pre-stimulus advice relative to post-stimulus advice (x-axis). Most datapoints fall in the lower left quadrant, indicating overall preference for post-stimulus advice. There were also some evidence of individual differences in this preference, such that participants who were more influenced by pre-stimulus advice tended to choose to receive pre-stimulus advice more.

Summary of key findings

1a) Characterise algorithm aversion in our task paradigm

- Overall we found no evidence of algorithm aversion in our perceptual decision task. Instead, the trend in the data, which was significant in some individual experiments, was for algorithmic advice to be more influential than human advice.
- The preference for algorithmic advice appears at least partly task-dependent: In the same sample of participants showing greater influence of algorithmic over human advice, we found evidence of algorithm aversion expressed in participants' preference for human decision making (e.g., in decisions about which individuals should receive vaccines during a flu outbreak).

1b) Investigate determinants of preferences for human vs. algorithmic advice

- Early experience of advice accuracy has a disproportionate impact on long-term preferences of advice and advisor.
- We observed large individual differences in preference for human vs. algorithmic advice.
- Individual differences in preference for human vs. algorithmic advice show some stability across tasks and can be captured in short survey tools, including our new Intelligent Systems Survey and Specific Systems Survey tools.

2) Identify factors promoting trust in algorithmic advice

- People expressed an overall preference for algorithmic advice provided with a graded estimate of advice reliability (confidence).
- Algorithmic advice was more influential and chosen more often when it followed vs. preceded people's own opportunity to view stimulus evidence.

Other achievements

Publications arising from the project to date are as follows:

Celaya A, Yeung N (2019). Human psychology and intelligent machines. In: A Gilli (Ed) The Brain and the Processor: Unpacking the Challenges of Human-Machine Interaction. NATO Defence College Research Paper, Chapter 2, pp. 17-26.

Celaya A, Yeung N (2019). Confidence and trust in human-machine teaming. Journal of the Homeland Defense & Security Information Analysis Center, 6(3), 21-25.

Celaya A, Aiyer S (2020). A Foundation of Automation for Future Artificial Intelligence Strategy. Journal of the Homeland Defense & Security Information Analysis Center, 7(1), 26-34.

Aiyer S, Celaya A, Jaquiery M, Yeung N (in prep). The Intelligent Systems Survey: a tool for assessing trust in automation and artificial intelligence. Manuscript in preparation for submission to *Human Factors*.

Dissemination

Publications arising from the project are listed above. In addition, research from the project was presented at the following meetings:

- NATO Defence College (Rome, Italy; December 2018): invited conference on *The Future of Warfare - Autonomous Systems*. Yeung and Celaya presented.
- SMi group meeting on Military Robotics and Autonomous Systems (London, UK; April 2019). Celaya presented.
- Changing Character of War Centre, University of Oxford. Workshop on How technology will affect the battlespace environment. Oxford, UK. May, 2019. Yeung presented.
- US Air Force, 711th Human Performance Wing Chief Scientist Lecture Series, May, 2021. Yeung presented.

Additional dissemination opportunities pursued during the project were:

- Yeung and Aiyer met with staff and researchers from the US Space Force and Air Force, including Dr Joel Mozer (Chief Scientist at USSF) and Dr Rajesh Naik (Chief Scientist, 711th Human Performance Wing, Air Force Research Laboratory).
- Celaya attended the following meetings as a delegate: Defence IQ meeting on Maritime ISR (Rome, Italy; September 2018); SMi group meeting on Military Space

Situational Awareness (London, UK; April 2019); Defence IQ Conferences on Land ISR & C2 Battle Management, and Space Operations, 28- 30 May 2019.
- Aiyer attended the following meeting as a delegate: Responsible Human-Machine Teaming workshop organised by the Alan Turing Institute. London, UK. October, 2019.

An important further avenue for dissemination of the research is via Aaron Celaya's military roles. After leaving Oxford in October 2019, he has continued to attend our project meetings virtually and has contributed to the ongoing work. During that time he has been stationed at Headquarters, United States Space Force (USSF), where he has served as the Artificial Intelligence Liaison for USSF to AFWIC at the Pentagon and served as the Deputy Branch Chief for USSF Doctrine and Concepts. Additionally, he is a Subject Matter Expert in the Space Trusted Autonomy collaboration between the USSF, NASA, Air Force Research Laboratory, and other agencies.

In parallel with the EOARD project, Yeung and Aiyer are part of a research consortium with joint US DoD / UK MoD funding, which aims to develop new frameworks for artificial intelligence agents to more truly team with human counterparts. This project continues to provide very valuable synergies with the EOARD project. For example, there has been considerable interest from that project in the survey tools we have developed here.

Impacts

Development of the principal discipline(s) of the project
Our project investigated the factors affecting trust in automated and intelligent systems. Contrary to previous evidence that people systematically mis-trust the outputs of algorithmic systems, we found little evidence of "algorithm aversion". Instead, the most striking feature of our experiments was the wide variation across individuals: Even when we carefully control the relative quality of human vs. algorithmic advice, we find wide variation in people's preferences for human vs. algorithmic advice. We find that some of this variability can be captured with simple survey tools to assess people's attitudes to artificial intelligence systems. Furthermore, we find that people are sensitive to subtle features of algorithmic advice, for example preferring algorithms that provide information about the likely reliability of their outputs.

Other disciplines
We have developed new survey tools assessing individual variation in trust in AI and automation that are relevant to human factors research. These tools have attracted interest from military stakeholders at our presentations, with ongoing discussions about future collaboration.

Describe the impact in this reporting period on the development of human resources
- Sriraj Aiyer initially worked on the project as a graduate Research Assistant. His experience on the project was vital in his successful application for a PhD position here in the Department of Experimental Psychology at the University of Oxford, which he began in October 2021.
- Lt Col Celaya gained valuable research experience and insight into effective human-machine teaming, which he has brought back to his station in the US Space Force, where he is currently a Lt Col with responsibilities including serving as the Artificial

Intelligence Liaison for USSF to AFWIC at the Pentagon and served as the Deputy Branch Chief for USSF Doctrine and Concepts.

Changes

The original proposed research included a component of human brain imaging with EEG and fMRI. However, with disruptions to in-person testing due to the pandemic, our attention focused on other components of the project, in particular in characterizing individual differences in trust in algorithmic systems.

Technical Updates

The project is now complete, with no further technical updates to report.