Scientific Reports

https://doi.org/10.1038/s41598-026-34983-y

Article in Press

Examining human reliance on artificial intelligence in decision making

Received: 31 March 2025

Accepted: 1 January 2026

Published online: 05 February 2026

Cite this article as: Pearson J., Dror I.E., Jayes E. *et al.* Examining human reliance on artificial intelligence in decision making. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-34983-y

Joe Pearson, Itiel E. Dror, Emma Jayes, Grace-Rose Whordley, Georgina Mason & Sophie Nightingale

Examining Human Reliance on Artificial Intelligence in Decision Making

Joe Pearson[1], Itiel Dror[2], Emma Jayes[3], Grace-Rose Whordley[3], Georgina Mason[3], and Sophie Nightingale[1]*

[1]Lancaster University, Department of Psychology, Lancaster, LA1 4YF, United Kingdom

[2] Cognitive Consultants International (CCI-HQ)

[3] Defence Science and Technology Laboratory

*s.nightingale1@lancaster.ac.uk

Abstract

The use of Artificial Intelligence (AI) to effectively support human decision making depends on whether humans are willing to rely on, and trust in, AI. Understanding human reliance on AI is critical given controversial reports of AI inaccuracy and bias. Furthermore, the erroneous belief that using technology removes biases may lead to overreliance on AI. To examine humans' reliance on AI, human participants (N = 295, M_{age} = 33.79) judged the authenticity of 80 faces (40 real, 40 AI-synthesized) presented alongside guidance supposedly from humans or from AI. This guidance was correct only half of the time. Participants indicated their confidence in each judgement and completed measures to examine propensity to trust humans and general attitudes towards AI. Participants who received AI guidance and exhibited more positive attitudes towards AI showed poorer discriminability between real and synthetic faces than those with less positive attitudes towards AI. For participants who received human guidance, level of trust in humans did not affect discriminability. Therefore, AI-derived guidance may be uniquely placed to engender biases in humans, leading to less effective decision making. To ensure successful human-AI decision making partnerships, more research is needed to understand precisely how humans use AI guidance in various contexts.

Keywords: computational social science, AI, decision-making, bias.

Introduction

Advances in technology and access to "big data" have allowed for the expansion of Artificial Intelligence (AI) to help with human decision-making. There are, potentially, significant benefits to using AI to support human decision-making, including saving time, improving accuracy, and reducing bias. Indeed, one need only consider the use of AI by various music and film streaming platforms and dating sites to grasp how accepted and *expected* the use of such technology has become [1, 2]. However, in recent years, the use of AI in certain contexts has led to much controversy, with reports indicating inaccuracy and unfairness; notably showing biases against certain demographic groups [3]. For example, the COMPAS tool used in many US states to predict a defendant's risk of recidivism has been

shown to be biased against people of colour [4]. Although there is a growing literature addressing the technical limitations of AI and trying to improve accuracy and fairness, there is insufficient research examining human-AI interactions. Specifically, the impact that guidance provided by AI, and a human decision-maker's general attitudes towards AI, will have on the accuracy and manifestation of biases in the individual's decisions.

The human tendency to over rely on technology is not a novel phenomenon. Over-trusting or over-relying on automated systems, including AI, is broadly defined as *automation bias* - a cognitive phenomenon whereby human operators tend to favour recommendations derived from a technological source over their own judgements [5]. One of the earliest critiques of automation came in the context of aviation where questions were raised around the safety of shifting from manual to automated flight-desk functions [6]. Even nearly half a century ago, the pace at which automation in aviation was advancing was described as outstripping the ability to comprehend the consequences for the different demands this change placed on human operators [7, 8]. Namely, interacting with automated systems draws on human operators in a different way, shifting demands on human cognition—from active control to more passive monitoring, where the human is expected to intervene in instances that the technology cannot handle [9]. Various aspects of human nature, human understanding of automated systems, and automation optimisation can contribute to the manifestation of automation bias. For example, much research has shown that when supervising automated systems, humans tend to demonstrate a loss of situational awareness which can lead to a failure to intervene to correct any errors (e.g., [10, 11]). Although much work has focused on the aviation sector, concerns about automation bias extend to other domains such as healthcare and military, especially with the relatively recent move to embed AI systems within these areas to assist human decision making [12, 13]. Automation bias has long presented a challenge in traditional computer-assisted decision-making and now poses critical concerns in the face of human-AI interaction.

Research has begun to reveal the extent to which human cognitive limitations restrict human-AI interactions and negatively impact real-world decision-making. People can, for example, succumb to a fallacy known as *technological protection* – the notion that the use of technology will remove biases [14]. This misplaced belief in the impartiality of technology may result in over reliance and misplaced trust in guidance derived from AI [15]. AI frequently provide outputs without the uncertainty cues common to human interaction. Response delays, disfluencies, and rephrasing during human interactions allow humans to gauge the credibility of incoming information. Without these cues, humans encountering AI may mistakenly attribute high confidence and, therefore, trustworthiness to AI outputs [16]. AI are also often portrayed as accessing and utilising all human knowledge [17]. Humans tend to accept, and form stronger beliefs based on, incoming information believed to be from a more credible and knowledgeable source [18, 19]. Furthermore, the widespread willingness to discuss and describe AI as anthropomorphic [20] reflects the known tendency of humans to readily assign human characteristics such as intentionality and—consequently—certainty to AI [21]. Thus, the design and portrayal of AI works with human nature, such that each has the potential to facilitate automation bias and, therefore, drive biased use of AI by human operators.

Importantly, greater reliance on and/or trust in automated technologies engendered by the aspects of human nature and AI design described above has long been understood to

increase use and misuse of such technologies by human operators [22, 23, 24, 25]. Indeed, frequent use and, therefore, familiarity with AI that is designed to support humans during decision-making tasks, may lead to problematic over-reliance on AI such that it is treated as an entirely autonomous decision-maker. One such example is the crash of Continental Flight 3407 in February 2009, wherein increased automation of cockpit procedures led to failures among crew to pay attention [26]. The consequences of human reliance on, and misplaced trust in, biased AI can also be seen in the use of Automated Fingerprint Identification Systems (AIFS). Human operators tend to over-rely on these systems to provide the most likely match at the top of the candidate ranking list, hence they make more false positive decisions on candidates on the top of the list and, conversely, more false negative decisions on candidates further down the list [27, 28].

Currently there is a lack of research and understanding about how humans and AI interact in decision-making tasks – do humans use AI effectively, or do they place too much reliance in the guidance AI provides, thereby reinforcing and legitimising cognitive biases within decision making? Importantly, to isolate the specific effect of AI, the research reported here included both AI and human guidance to allow comparison of results across these input types. An understanding of the effect of AI guidance on decision-making is important to ensure maximal benefit from technological advances while avoiding pitfalls. This research will highlight potential ways to use AI more effectively, for example how to reduce bias and under what circumstances are biases most likely. As such, beginning to understand the characteristics of AI-created bias will allow a more informed appreciation of how AI can benefit strategic and operational planning, while a greater understanding of perceptions and trustworthiness of AI in the decision making process will increase transparency. Overall, understanding the impact of AI on human cognitive bias is crucial before AI is widely deployed and integrated into human decision-making procedures. The aim of the current research is to provide the initial steps in gaining such an appreciation.

To do so, this research examined whether humans use AI and human guidance in a useful way – that is, rely on the guidance if it is accurate and dismiss it if it is inaccurate. We used a relatively simple decision-making task – determining whether a face is real or synthetic [29]. There are two clear benefits to using this task: 1) the stimuli set is already validated and 2) previous research provides baseline accuracies allowing us to select facial images that, although difficult to classify, can be accurately classified by most people (accuracy range of 64-84%). The study employed a mixed experimental design: a between-subjects manipulation wherein participants received either human or AI guidance; within-subjects manipulations of stimulus authenticity (real vs synthetic faces) and guidance accuracy (correct vs incorrect).

Research questions

1. Will participants who receive guidance (from AI or humans) show a similar decision accuracy to a baseline control comparison group (from [29])?
2. How will response accuracy (correct identification of faces as either real or synthetic) be affected when participants are given incorrect guidance vs. correct guidance?
3. Will participants who receive AI guidance provide more responses consistent with that guidance than participants who receive human guidance?

Method

Participants

Participants were recruited via Prolific, an online participant recruitment platform with over 130,000 members vetted to take part in research studies. Prolific users were eligible for participation if they: reported themselves to be fluent in English, have >95% approval rating, use a desktop computer/laptop with screen size >1024x768 pixels, ≥18 years-of-age, and have normal/corrected-to-normal colour vision. A sensitivity power analysis appropriate for two-way ANOVA showed that a sample size of 274 yields a power of .80 for a small-to-medium effect size of .17 and an α of .05. Data were collected from 322 Prolific users (26 participants' data were removed due to: device/operating system check = 9, vision check = 2, attention checks = 9, guidance use check = 2, withdrawn consent = 3, non-complete = 1). Following data cleaning, a final sample of 295 individuals (M_{age} = 33.79, SD = 10.76) remained. Of these 295 participants, 182 identified as male, 109 as female, 2 as non-binary/genderqueer/agender/gender fluid, 1 as transgender male, while 1 preferred not to say. Additionally, 209 self-reported as White, 58 Black, 12 Mixed, 11 Asian, and 5 Other. Participants were paid £5 upon completion of the experiment, and the amount paid did not depend on a participant's performance in the face classification task.

Materials

Stimuli. The real and synthetic faces used in this study were taken from Nightingale and Farid (2022; [29]), where a stimulus set of 400 real and 400 synthetic faces was created. Thus, the stimuli used here have been previously validated and have baseline accuracies, allowing us to select stimuli that fall within a certain mean accuracy range (64-84%) that, although difficult to make judgements about, can be accurately categorised as real or synthetic by most people. This accuracy range yielded an available stimulus set of 156 faces (102 real, 54 synthetic). For the current study, 80 stimuli (40 real, 40 synthetic) were selected.

Table 1

Number of real and synthetic faces of each available gender and ethnicity.

Gender/ethnicity	Real		Synthetic	
	Count	Mean (SD)	Count	Mean (SD)
Male	20	.73 (.06)	19	.73 (.06)
Female	20	.72 (.05)	21	.71 (.06)
Black	11	.72 (.06)	16	.72 (.06)
East Asian	6	.72 (.06)	14	.71 (.06)
South Asian	12	.73 (.06)	9	.73 (.05)
White	11	.72 (.06)	1	NA
		.71 (.06)		**.71 (.06)**

Note. Mean and standard deviation (SD) accuracy data derived from [17].

Whereas stimuli were selected to represent a diverse population, the number of real and synthetic faces available within the prescribed accuracy range made equal representation

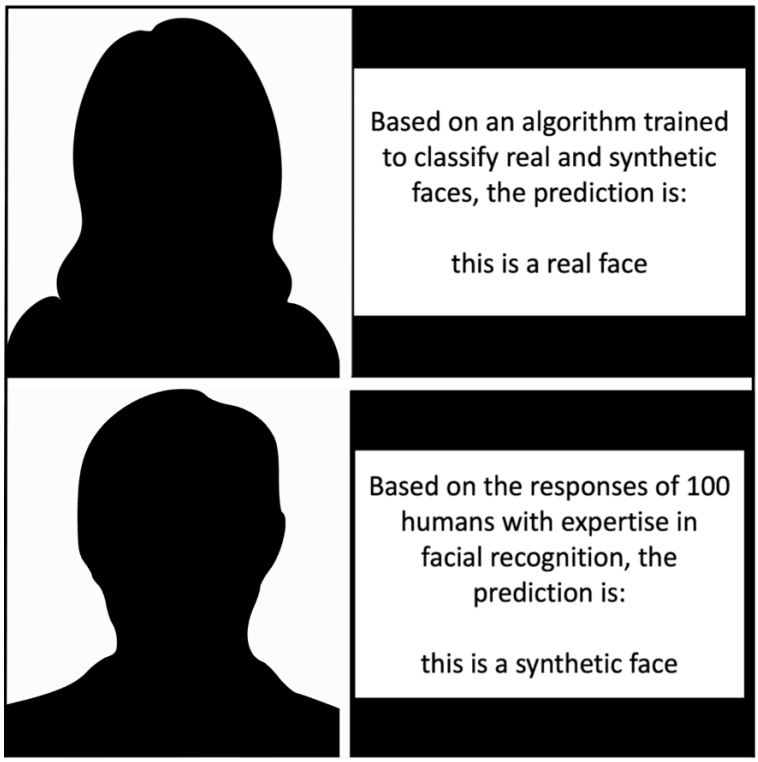
of gender (male, female) and ethnicity (Black, East Asian, South Asian, White) impossible. The distribution of real and synthetic stimuli across gender and ethnicity identifiers was completed as evenly as possible according to the first author's discretion (Table 1). The full stimuli selection protocol is available in supplementary materials (SM) 1.

Guidance Information. Each face was paired with correct or incorrect human or AI guidance information, producing standardised information cards (Figure 1). In both the human and AI guidance conditions, half of the cards provided correct guidance information (e.g., stated a real face was real) and half provided incorrect information (e.g., stated a real face was synthetic). In both the human and AI conditions, 'A' and 'B' streams were constructed to counterbalance the appearance of stimuli alongside correct or incorrect information (i.e., if a face appeared alongside correct guidance in A, it appeared alongside incorrect guidance in B). All guidance streams included all 80 faces, while 20 of each type of face (real or synthetic) were presented alongside correct guidance and 20 alongside incorrect guidance. Stimuli appeared in a random order, and participants were unaware of the manipulation of real vs. synthetic faces and correct vs. incorrect guidance.

Face classification task. Participants responded to all 80 faces. In each trial participants indicated if the face was of a real person or if it had been artificially synthesized. They reported their level of confidence in each judgement using a Likert scale (1 = not at all confident, 5 = extremely confident).

Figure 1

Example stimuli with facial images shown as silhouettes due to licencing permissions: top = synthetic face, AI condition; bottom = real face, human condition. Real faces were obtained from Flikr-Faces-HQ Dataset [30], made available by NVIDIA Corporation under Creative Commons BY-NC-SA 4.0 license.



Note. 'Prediction' is favoured over alternatives (e.g., 'conclusion') since it suggested an estimate that cued participants to evaluate the stimuli before providing a response. Due to licensing permissions, facial images are shown here as silhouettes to illustrate where the real/synthetic faces appeared.

Attention check stimuli. To ensure participants engaged with the face classification task, four attention check trials were presented in the first two thirds of the study. These were synthetic faces containing various errors such as missing, misshapen, or discoloured features. To ensure participants knew what these attention check images might look like, they were given a description of the types of errors they could reasonably expect and were also shown four examples of poorly synthesised faces. Participants were informed that these erroneous images constituted attention checks following the same presentation format as the experimental stimuli, and were screened out from the study if they failed to correctly identify at least three of the four attention checks.

Human trust scale. Participants completed the human trust scale (SM 2), a 17-item questionnaire drawing on items from several other scales [31, 32, 33, 34, 35, 36] to measure participant beliefs about others' honesty and trustworthiness. Items are scored from 1 (Strongly Disagree) to 5 (Strongly Agree), and items 7, 8, 9, 10, 11, 12, and 13 are reverse-coded. The score for each item is averaged together to form a continuous measure of generalised trust, such that higher scores indicate greater trust in humans. Example items include: "Most people are basically honest", "Most people are trustworthy", and "I usually trust people until they give me a reason not to trust them". Confirmatory factor analysis (CFA) was used to determine the construct validity of this composite measure of human trust. Internal consistency for the human trust scale used in data analysis was excellent (Cronbach's $\alpha = .89$, 95% CI [.87, .91]; [37]).

General Attitudes towards Artificial Intelligence Scale (GAAIS). Participants also completed the GAAIS [38], a validated 20-item scale assessing attitudes towards AI across positive and negative subscales. Positive subscale items – e.g. "Artificially intelligent systems can help people feel happier" – and negative subscale items – e.g. "I find Artificial Intelligence sinister" – are scored from 1(Strongly disagree) to 5 (Strongly agree), but negative subscale items (3, 6, 8, 9, 10, 15, 19, 20) are reverse-coded. Separate overall scores for the positive and negative subscales are computed by taking the mean score of each set of items. The higher the score on each subscale, the more positive the attitude toward AI. Both the positive and negative subscales demonstrated excellent (α = .91, 95% CI [.89, .92]) and good (α =.85, 95% CI [.83, .88]) internal consistency respectively.

Guidance use check. Participants completed a final survey item examining how much they used the guidance to inform their judgements on the face classification task. Participants indicated whether they 1) read the guidance information and always used it to help them decide if each face was real or synthetic, 2) sometimes used it, 3) read the guidance information but did not use it, or 4) did not read the guidance information. Individuals who reported not having read the guidance were excluded from data analysis (n = 2).

Procedure

This study was constructed and completed using Qualtrics and published online via Prolific. Before taking part, participants were informed of the study's purpose, their right to withdraw and payment details, and researcher contact information. Informed consent was obtained electronically, after which participants were subject to a device- and vision-check and advised that use of an ineligible device would result in non-payment. Before the experimental task began, participants saw several example stimuli and completed three practice trials to ensure they understood the task. The practice trials included a real, synthetic, and attention check face, and participants received feedback on each of their responses. Participants were then informed they would see 84 faces alongside guidance that may be of use, although they were not informed that 1) the guidance had been fabricated for the purposes of the study, 2) they had been randomly assigned to one of two experimental conditions, or 3) the presented stimuli had been deliberately balanced.

Participants were randomly assigned to one of the four counterbalanced guidance streams, wherein they completed the experimental task outlined above. Following the experimental task, participants also completed the human trust scale and GAAIS. Finally, participants provided age, gender, ethnicity, and guidance use information. Upon completion participants were fully debriefed. The previously undisclosed experimental manipulations were made clear, as were the true aims of the research. Participants were also reminded of their right to withdraw their data from the study and given the option to do so.

Analyses

Independent samples t-tests were conducted to determine differences in response accuracy and consistency between human and AI guidance conditions. A one-sample t-test was performed to compare accuracy scores for both human and AI guidance groups with a baseline control group [29]. One- and two-way ANOVAs were carried out to identify differences in response accuracy and consistency at different levels of self-reported guidance use within and across guidance conditions. For these analyses, accuracy scores reflect the

proportion of experimental trials a participant classified correctly, while consistency scores reflect the number of trials in which a participant responded in line with the guidance provided (regardless of accuracy).

Signal detection analyses assessed classification accuracy and response bias using indices of discriminability and criterion shift (d' and c respectively, [39, 40]). Independent samples t-tests on $d'$and c values compared face classification task performance between guidance conditions. Linear regression analyses assessed the influence of human trust scale and GAAIS positive and negative subscale scores on d' and c.

Ethical approval was granted by Lancaster University's Faculty of Science and Technology Research Ethics Committee (FST-2023-3241-RECR-3) and Ministry of Defence Research Ethics Committee (2213/MODREC/23), and this experiment was performed in accordance with relevant guidelines and regulations. These methods and planned analyses were preregistered at https://doi.org/10.17605/OSF.IO/SRTHP. Several additional exploratory analyses not preregistered were conducted and have been identified below. Data tidying was completed using R (Version 1.4.1106, RStudio Team, 2021) and Microsoft Excel, and all analyses were completed in R.

Results

Independent samples t-tests indicated no significant differences in response accuracy between counterbalanced human and AI guidance streams. Consequently, both sets of A and B guidance streams were collapsed into one human and one AI guidance group (see SM 3 for preliminary visualisations (Figure S1) and analyses).

Response accuracy and consistency. Table 2 shows mean response consistency for each guidance group. Individuals were more inclined to make judgements consistent with the guidance provided when it correctly classified faces as real or synthetic we did not find differences in guidance use across demographic characteristics (SM 4 Table S1). A paired-samples t-test was conducted to compare mean consistency scores for correct and incorrect guidance information, revealing a significant difference between the two ($t(294) = 29.74$, 95% CI = [12.17, 13.89], $p < .001$). The effect size (Cohen's d, [41]) of 1.73 indicates a large effect. It seems reasonable to conclude, therefore, that participants used the guidance strategically, relying on it more often when it was useful but disregarding it more frequently when it was not.

Table 2

Mean and standard error (SE) consistency scores (counts) for stimuli with correct and incorrect guidance, and all stimuli, across guidance conditions.

Guidance group	Correct (out of 40)	Incorrect (out of 40)	Correct & Incorrect (out of 80)
	Mean (SE)		
AI	30.50 (.36)	17.90 (.58)	51.55 (.86)
Human	30.20 (.35)	16.70 (.51)	50.12 (.75)
	30.30 (.25)	**17.30 (.38)**	**50.80 (.57)**

A Mann-Whitney-Wilcox test, appropriate for non-normally distributed data, indicated no significant difference in response consistency between human and AI guidance groups (w = 11396, 95% CI = [-1.00, 3.00], p = .46; human mean = 50.12, AI mean = 51.55). An independent samples t-test revealed no significant difference in response accuracy between human and AI guidance (t(293) = -.99, 95% CI = -.03, .01], p = .32; human mean = .67, AI mean = .66). It appears that the accuracy with which individuals correctly classified real and synthetic faces, and the extent to which they classified such faces consistently with the guidance, did not change with the guidance source. Additionally, a one-sample t-test indicated no significant difference (t(294) = -1.28, 95% CI == [.65, .67], p = .20) between overall response accuracy (.66) and a baseline accuracy level of 67% [29].

A one-way between-subjects ANOVA revealed a significant effect of level of guidance use (*Always used* vs. *Sometimes used* vs. *Did not use*) on response accuracy (F(2, 292) = 11.71, p < .001). The effect size, as measured by generalised eta^2 (η_g^2), was .07 (small effect). Pairwise comparisons using the Tukey method (Table 3) revealed the *Always used* group's mean accuracy was significantly lower than the *Did not use* (coefficient estimate = .08, 95% CI = [.04, .12], p < .001), and *Sometimes used* groups (.05, 95% CI = [.02, .08], p < .001). A one-way ANOVA examining the effect of guidance use level on response consistency revealed a significant effect of guidance use level (F(2, 292) = 9.35, p < .001, η_g^2 = .06,). Pairwise comparisons using the Tukey method (Table 3) revealed that the *Always used* group mean consistency was significantly greater than the *Did not use* group (-7.48, 95% CI = [-11.80, -3.20]), as was the *Sometimes used* group (5.97, 95% CI [2.27, 9.68], p < .001).

Table 3

Mean and standard error (SE) response accuracy (%) and consistency scores (counts) across guidance use levels.

Guidance use level	Accuracy	Consistency		
	All stimuli		Correct guidance	Incorrect guidance
	Mean (SE)			
Always used	.62 (.01)	52.90 (1.42)	29.60 (.64)	20.10 (.90)
Sometimes used	.67 (.01)	51.39 (.68)	30.80 (.30)	17.10 (.46)
Did not use	.70 (.02)	45.41 (1.04)	29.60 (.54)	14.00 (.76)

Individuals who reported using the guidance information at their own discretion or not at all performed better than those who claimed to have always used it. This result is not surprising given that adherence to all guidance would yield just 50% accuracy. Of particular interest, though, is the discrepancy between the mean total response consistency of those individuals reporting to have always used the guidance, and the expected consistency of this group. Always using the guidance should yield a consistency score of 80, since judgements made by these individuals are made in line with the available guidance regardless of whether it is correct. It seems, therefore, that some participants misremembered their reliance on the

guidance. Visualisations (SM 5, Figure S2) and analyses (SM 5) examining the effect of the interaction between guidance use level and guidance stream and response accuracy and consistency revealed non-significant effects on both response accuracy ($F(2, 289) = .82$, $p = .44$) and consistency scores ($F(2, 289) = .76$, $p = .47$).

Signal Detection Analyses. d' and c (computed in R using the *psycho* package, [42]) represent sensitivity to the difference between real and synthetic faces, and criterion shift (inclination to respond more in one direction than another). These values are derived from counts of hits (correct classification of a face when guidance is correct), correct rejections (correct classification when the guidance is incorrect), misses (incorrect classification when the guidance is correct), and false alarms (incorrect classification when the guidance is incorrect, [43]).

Table 4 shows mean d' and c scores for human and AI guidance groups and for the entire dataset. A d' value of zero indicates no ability to distinguish between real and synthetic faces, and a value of 3 represents close to perfect discrimination. For c, a value of zero indicates no response bias (equally likely to respond 'real' or 'synthetic'), negative values indicate that an individual responds 'real' more often, and positive values indicate that an individual responds 'synthetic' more often. Mean d' (.90, 95% CI [.84, .97]) and c (-.28, 95% CI [-.33, -.24]) values suggest participants showed an ability to distinguish between real and synthetic faces but a tendency toward responses of 'real'. To determine if d' and c values are significantly different to 0, one-sample Wilcox t-tests were carried out. These tests revealed that at the dataset level ($v = 41986$, 95% CI = [-.15, .10], $p < .001$, Cohen's $d = 1.62$) and in each guidance group d' was significantly above 0 (Human: $v = 11430$, 95% CI = [.83, 1.01], $p < .001$, $d = 1.52$; AI: $v = 9682.50$, 95% CI = [.81, .96], $p < .001$, $d = 1.80$). The same was observed for a series of one-sample Wilcox t-tests performed on c data at the dataset ($v = 1813.50$, 95% CI = [-.27, -.21], $p < .001$, Cohen's $d = -.81$), Human ($v = 593$, 95% CI = [-.26, -.19], $p < .001$, $d = -.83$) and AI ($v = 334$, 95% CI = [-.31, -.21], $p < .001$, $d = -.81$) levels. Participants displayed a significantly better than chance ability to distinguish between real and synthetic faces, but a significant bias toward identifying faces as 'real'. These findings are supported by the positive skew in d' distribution and substantial negative c distribution illustrated in Figure S3 (SM 6).

Table 4

Mean and 95% confidence intervals (CI) of d'and c values for each guidance stream.

Guidance stream	d'	c
	Mean (95% CI)	
Human	.94 (.84, 1.03)	-.27 (-.31, -.21)
AI	.87 (.79, .95)	-.32 (-.38, -.25)
	.90 (.84, .97)	**-.28 (-.33, -.24)**

Mann-Whitney-Wilcox tests appropriate for non-normally distributed data revealed no significant difference in d' ($w = 10564$, 95% CI = [-.15, .10], $p = .69$) or c ($w = 10256$, 95% CI = [-.08, .03], $p = .41$) scores between guidance groups. It appears, therefore, that the type of guidance an individual receives when making judgements about the nature of real or

synthetic faces influences neither their ability to distinguish between the two nor their bias in responding.

To determine if the composite questionnaire used here to examine trust in other humans assesses a latent construct of trust, a confirmatory factor analysis (CFA) using maximum likelihood estimation was conducted (SM 6, Table S2). The model specified one latent variable (trust) underlying all observed indicators (excluding item 11, an attention check). Model fit was determined by examining: Chi-square (X^2), a measure of overall model fit; Root Mean Square Error of Approximation (RMSEA) and Standardised Root Mean Square Residual (SRMR), measures of how far a model is from perfect fit; Tucker-Lewis Index (TLI) and Comparative Fit Index (CFI), which compare model fit to the worst possible model. The model demonstrated poor fit to the data, as indicated by a significant X^2 test ($X^2(119) = 967.20$, $p < .001$). TLI and CFI scores of .69 and .73 fall below the commonly accepted threshold of .90 for adequate fit [44], while RMSEA and SRMR values of .16 and .09 exceed the typical cutoff scores of .08. Together, these results indicate that a one-factor structure does not adequately represent the data.

To address the poor fit of this unidimensional model of trust in humans, only those items assessing 'propensity to trust' were taken forward to analysis. These four items – 'I usually trust people until they give me a reason not to trust them', 'trusting another person is not difficult for me', 'My typical approach is to trust new acquaintances until they prove I should not trust them', and 'My tendency to trust others is high' – were taken from [31] and constitute a validated measure of propensity to trust other humans. A confirmatory factor analysis (CFA) using maximum likelihood estimation revealed an acceptable fit for this model. All item loadings were significant (higher than 0.82) and fit statistics were good CFI = .99, TLI = .96, and SRMR = .02, aside from chi square ($\chi 2$ (2) = 10.93, p = .004) and RMSEA = 0.12 [90 % CI 0.06 – 0.20]. Internal consistency for this new human trust scale was excellent (Cronbach's α = .89, 95% CI [.87, .91]; [44]). The average inter-item correlation was .67, indicating strong internal consistency among the items. Following recommendations to report and consider the model indices in combination [45, 46, 47], these results suggest that these four items constitute a reasonable measure of propensity to trust other humans. A single measure of human trust was created in accordance with [31]'s recommendations by taking the mean score across the four propensity to trust items. Internal consistency was assessed for both the positive and negative GAAIS subscales. The positive and negative subscales demonstrated excellent (α = .91, 95% CI [.89, .92]) and good (α =.85, 95% CI [.83, .88]) internal consistency respectively. The items comprising each subscale measure a common construct. The full CFA process is described in SM 6.

To determine if general attitudes towards AI or propensity to trust humans influences task performance, linear regressions were conducted with d' and c as dependent variables and GAAIS subscales and propensity to trust other humans scale scores as independent variables. A significant effect of the GAAIS negative subscale score on d' was identified (b = -.15, SE = .05, p = .004). Thus, more positive attitudes toward AI yielded a reduced ability to discriminate between real and synthetic faces. Interestingly, the effect of greater positive attitudes towards AI on discriminability was preserved when the same regression model was fit using AI guidance group data only (b = -.19, SE = .07 p = .008), but not when fit using the human guidance group data only. A significant effect of human trust scale on c values was observed (b = -.05, SE = .02, p = .03), such that a greater propensity to trust other humans

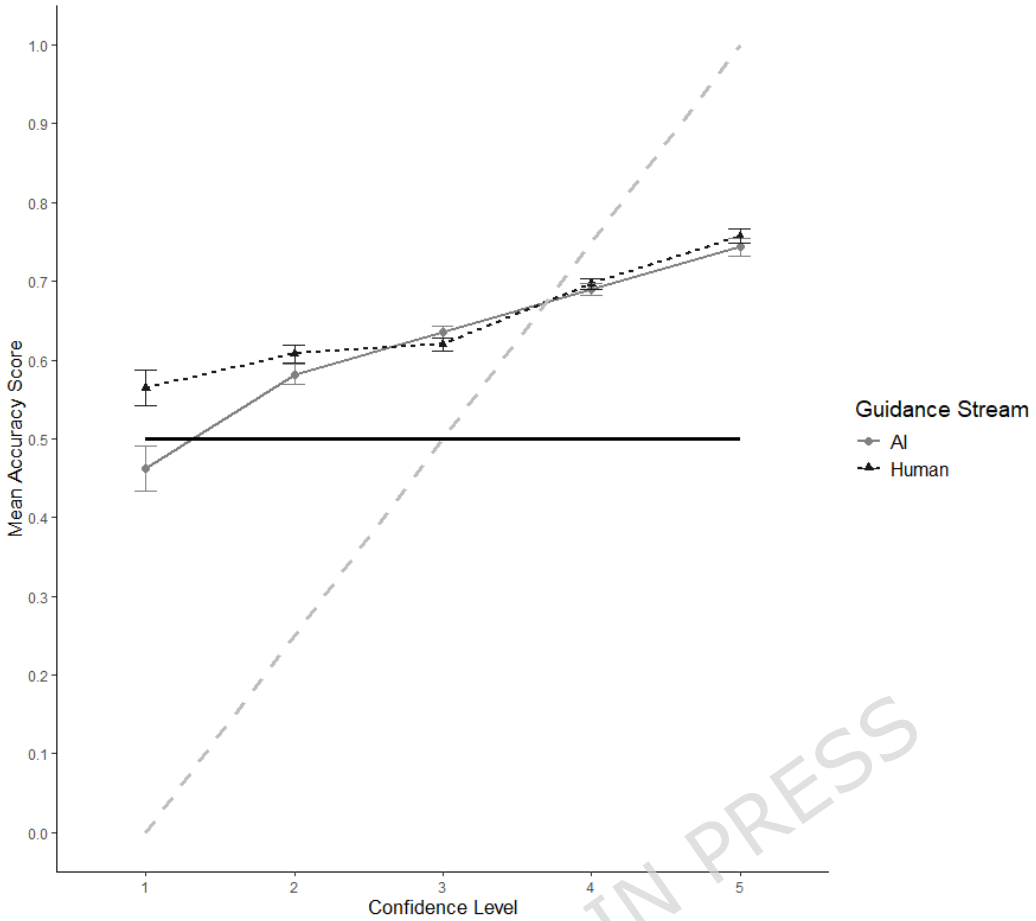
predicted a shift toward face classifications of 'real'. Exploratory analyses were carried out to control for the effects of level of guidance use, guidance stream, participant age, gender, and ethnicity. These parameters were entered into each regression model sequentially, and the impact of increased model complexity on model fit was assessed using ANOVA tests. The full model-building process is described in SM 6.

A linear regression with d' as dependent variable and GAAIS positive and negative subscale scores, human trust scale score, and self-reported level of guidance use (*Always used*, *Sometimes used*, *Did not use*) as independent variables was performed. The previously identified effect of GAAIS negative subscale on task performance was preserved (b = -.13, SE = .05, p = .008). No effect of GAAIS positive subscale or human trust scale scores were observed. A significant effect of self-reported guidance use was observed, so that participants who did not use the guidance showed significantly larger d' scores than those who always used it (b = .38 SE = .10, p < .001), as did those who only sometimes used it (b = .25, SE = .08, p = .001). Lower self-reported guidance use predicted an improved ability to discriminate between real and synthetic faces. The same regression model was fit with c score as the dependent variable. The previously identified effect of human trust scale on c values was preserved (b = -.05, SE = .02, p = .009). Additionally, a significant effect of self-reported guidance use level on c value was observed – participants who did not use the guidance showed significantly larger c values than those who always used it (b = .26, SE = .07, p < .001). Thus, less reliance on guidance predicted a reduced likelihood of classifying faces as real. ANOVA revealed a significant improvement in model fit (p < .001) with the inclusion of guidance use level for both sets of regression models.

A linear regression with d' as dependent variable and GAAIS positive and negative subscale scores, human trust scale scores, self-reported level of guidance use, and guidance stream as independent variables was performed. No significant effect of guidance stream on d' was identified, while the pattern of results for GAAIS subscales, human trust scale, and self-reported guidance use level remained identical to that observed in the previous model. Furthermore, ANOVA revealed a non-significant (p = .31) improvement in model fit with the inclusion of guidance stream. The same regression model was fit with c score as the dependent variable. No significant effect of guidance stream on c score was observed. The pattern of results for the remaining independent variables were identical to those identified in the previous modelling iteration. ANOVA revealed a non-significant (p = .16) improvement in model fit with the inclusion of guidance stream. Two final regression models accounting for the influence of participant sociodemographic characteristics on d' and c scores were created, revealing significant effects of participant gender on d' scores (b = .17, SE = .07, p = .01), and of age on d' (b = -.01, SE = .003, p < .001) and c (b = -.004, SE = .002, p = .04), and. It appears that women showed an increased ability to discriminate between real and synthetic faces, while older participants showed a decreased ability to discriminate between real and synthetic faces and a greater likelihood of classifying faces as real.

Figure 2

Confidence-accuracy curve of mean accuracy scores (and standard error bars) per level of confidence for AI and human guidance.



Note. Diagonal dashed grey line = perfect calibration. Full black line = chance (50% response accuracy).

Confidence-accuracy calibration. Confidence data collected at each trial (5-point Likert scale: 1=not at all confident, 5=extremely confident) were combined with response accuracy data to examine participants' insights into their decisions. The combined data show the relationship between response accuracy and confidence in judgements (Figure 2). Those participants who were more confident in their judgements performed better than those who were less confident, regardless of whether they received human or AI guidance.

Additionally, meta-d' values were calculated for each guidance condition. Meta-d' is a statistical counterpart to the confidence-accuracy relationship visualised above, measuring 'metacognitive sensitivity' [48, 49]. Mean meta-d' values of .09 (95% CI [.04, .18]) and .13 [.03, .18] were observed for the human and AI guidance groups respectively. Positive values occur when high levels of confidence are reported for correct judgements, and low levels for incorrect judgements. In line with the confidence-accuracy curves (Figure 2) it appears that participants in both guidance groups showed some ability to recognise whether they were making correct and incorrect judgements.

Discussion

This study examined how reliance upon AI guidance during decision-making influences human judgement and cognitive bias. Participants who received guidance from a

supposed human source were as accurate and consistent in their classifications of faces as real or synthetic as those who received AI guidance. Furthermore, a significant response bias towards face classifications of 'real' was observed for individuals who received AI guidance, while regression analyses identified reduced task performance (measured by d') for participants with more positive attitudes towards AI than those with less positive attitudes towards AI.

Task performance was not affected by the source of the guidance an individual received. Additionally, it was found that classification consistency *did not* differ significantly with guidance source. This latter finding is surprising given the known tendency for individuals to succumb to the fallacy of technological protection [14]. Indeed, we might have expected participants who received AI guidance to show increased consistency with this guidance regardless of whether it was correct or incorrect. Yet, participants showed a similar level of adherence to AI as to human guidance. Furthermore, participants were more inclined to follow guidance when it was correct regardless of its source. Humans appear able to follow guidance when it is correct and disregard it when it is not.

The apparent strategic use of AI guidance is encouraging given what we know about the tendency for over-reliance by humans on potentially biased AI [15] born out of misplaced positivity towards, and/or trust in, such systems. Thus, during human-AI decision-making interactions it seems that rather than AI protecting against biases – as the fallacy of technological protection might suggest [14] – it is human decision-makers that work to mitigate biases. In this regard, discrepancies between AI predictions, the subject of a decision (e.g., a real or synthetic face), the specific knowledge of human decision-makers [50], and prior experience and familiarity with AI systems and their outputs [21, 51] may prevent human operators from following ineffective AI support. Other research, however, highlights the human as the problematic component during human-AI decision-making interactions [52]; the researchers observed in a face-matching task that humans do not perform as well when supported by a simulated automated facial recognition system (sAFRS) as the same sAFRS by itself, due to overturning of correct sAFRS predictions but failures to overturn sAFRS mistakes.

If it is true that humans are a mitigating force against AI biases, one must ask whether it is worthwhile utilising AI in human decision-making at all. At the very least, our focus must be on developing human-AI decision-making interfaces that optimise the regulatory role of humans. Supportive AI of this nature has been of interest to the algorithmic fairness research community for some time [53], yet how they influence decision-making performance remains unclear [54, 55]. If humans are detrimental to human-AI decision-making interactions then the result is the same. Until the issue of AI bias can be resolved – which requires reforms to big data collection practices – it is the human component that we must depend on and enhance to ensure effective AI use.

The stability of response consistency across guidance conditions may also be informed by whether participants relied on the guidance. This seems likely given many participants reported using the guidance only some of the time, whether it was derived from humans or AI. Furthermore, no significant difference in response consistency was observed between those who reported having always used the guidance and those who used it only some of the time. It seems that participants used the guidance as and when they deemed it

necessary. This is a sensible and effective strategy highlighting future opportunities for human-AI decision-making partnerships. Occasions under which reliance on available guidance was necessary may have arisen when specific experimental trials presented a difficult choice, such as faces that appear quite but not entirely either real or synthetic. This seems reasonable given previous research identifying increased reliance on advice [56], and algorithmic advice [57], by humans when tasks are difficult. Furthermore, automated support system research has highlighted reduced trust in systems deployed on simple tasks [58]. The perceived difficulty of each trial may have determined participant guidance use and may be the mechanism underlying strategic guidance use. Participants likely will have relied on the guidance when they struggled with a decision and disregarded it when confident in their judgements.

Linear regression analyses revealed a significant effect of GAAIS negative subscale score on discriminability amongst individuals who received AI guidance. For individuals with more positive attitudes towards AI, decision-making effectiveness is reduced when they encounter AI guidance. Previous research identifying poorer decision-making amongst humans more frequently using and therefore trusting AIFS [27, 28] seems to support this observation, reaffirming the conclusion that the effectiveness of AI depends on the humans being supported, task difficulty, and guidance quality. This finding bolsters those of previous research identifying the importance of individual differences in trust in AI on human-AI decision-making partnership success, wherein large performance gains have been observed amongst humans re-completing a face-matching task with sAFRS support, especially when they held favourable beliefs about the system [59]. Under these circumstances, the importance of the human at the heart of such interactions is recorded in whether individual differences in trust in technology impedes their acceptance of AI support. That greater trust in humans did not influence discriminability similarly for those who received human guidance suggests AI may be uniquely placed to manifest changes in decision-making ability.

Interestingly, regression analyses showed a significant effect of propensity to trust other humans scale score but not GAAIS score on response bias. Among individuals reporting a greater propensity to trust other humans there is an increased likelihood to identify faces as real. Given that other analyses identified a tendency to classify faces as real regardless of the guidance source, it may be that in a face classification task of this kind participants' default position is that stimuli depict real faces. For individuals with greater trust in humans, classifications may default in this direction more readily. This is at odds with the previously discussed notions that humans use guidance strategically and that they can act as a regulatory force in human-AI interactions. Why this default position is not overcome and a bias toward classifications of faces as synthetic observed amongst individuals displaying greater positivity towards AI remains unclear. It is worthwhile noting that the CFA fit indices reported here for the propensity to trust other humans scale were mixed, some suggesting a good model fit and others suggesting a weak model fit. The use of this scale is theoretically-driven with it having been developed and validated by existing research [31], nonetheless, we have cautiously interpreted the human trust results to ensure that the conclusions of this work are valid and useful to the field.

Confidence-accuracy curves suggest that participants in both guidance conditions were able to reflect on their judgements effectively. Positive mean meta- d' scores for both guidance groups support this conclusion. That participants demonstrated a good

understanding of their capabilities extends similar observations from previous research utilising face stimuli [60] but contradicts other research identifying unjustified confidence during decision-making [61, 62]. The difficulty of this experimental task was controlled to ensure it was possible (by selecting stimuli between a previously identified classification accuracy bracket of 64-84%, [29]). It may be that previous research identifying poor participant insight employed more difficult tasks. This would explain the overall good performance and fair insight displayed by participants.

The various and differing findings observed here mean that more work is required to understand the circumstances under which AI biases manifest, and the role of the human in human-AI interactions. Indeed, given the importance of the human operator being supported in determining AI effectiveness, further investigation of the individual differences influencing the impact of AI on decision making should be prioritised. Future research should manipulate the previously discussed occasions of necessity under which guidance is utilised by humans during decision making, by using decision-making tasks of varying difficulty. This may be achieved by presenting both correct and incorrect guidance with varied accuracy information during a face classification task, yielding scenarios wherein ostensibly highly accurate predictions contradict the accompanying stimuli. Additionally, this experimental paradigm should be applied to various decision-making contexts. Human-AI interactions in, for example, critical military reconnaissance scenarios may manifest biases differently to those in low-demand online experiments.

By developing an increasingly nuanced conceptualisation of human-AI decision-making interactions, and the variation in these interactions across contexts, more effective AI and protocols for their use can be developed. It is imperative, though, that these tools are developed with the best interests of human operators in mind and deployed with fully informed human decision-makers at the heart. Ultimately, AI without human intervention can be useful, but our findings suggest that it is humans who decide how and when.

Data Availability

The data collected during this research, and the full, anonymised, reproducible R data tidying and analysis code is available at
https://osf.io/2p3bf/?view_only=868c92c940c947b894d24ac4b4155607

References

1. Bogert, E., Lauharatanahirun, N., & Schecter, A. Human preferences toward algorithmic advice in a word association task. *Scientific reports*, **12**(1), 14501; 10.1038/s41598-022-18638-2 (2022)
2. Hickey, A. *How Coffee Meets Bagel leverages data and AI for Love.* CIODIVE. https://www.ciodive.com/news/coffee-meets-bagel-dating-technology-ai-data/548395/#:~:text (2019).
3. Mauro., G., & Schellman, H. 'There is no standard': investigation finds AI algorithms objectify women's bodies. *The Guardian.* https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies (2023).
4. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. *Machine Bias*. ProPublica. Retrieved September 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
5. Cummings, M. L. Automation bias in intelligent time critical decision support systems. In: Collection of technical papers – AIAA 1st intelligent systems technical conference, **2**, 557 – 562; 10.2514/6.2004-6313 (2004).
6. Wiener, E. L., & Curry, R. E. Flight-deck automation: Promises and problems. *Ergonomics, 23*, 995-1011. https://doi.org/10.1080/00140138008924809 (1980).
7. Bainbridge, L. Ironies of automation. *Automatica, 19*, 775–779. https://doi.org/10.1016/0005-1098(83)90046-8 (1983).
8. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30*, 286-297. https://doi.org/10.1109/3468.844354 (2000).
9. Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human Factors*, *59*, 5-27. https://doi.org/10.1177/0018720816681350
10. Foroughi, C. K., Devlin, S., Pak, R., Brown, N. L., Sibley, C., & Coyne, J. T. Near-perfect automation: Investigating performance, trust, and visual attention allocation. *Human Factors, 65*, 546-561. https://doi.org/10.1177/0018720821103288 (2023).
11. Kaber, D. B., Onal, E., & Endsley, M. R. Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human factors and ergonomics in manufacturing & service industries, 10*, 409-430. https://doi.org/10.1002/1520-6564(200023)10:4<409::AID-HFM4>3.0.CO;2-V (2000).
12. Goddard, K., Roudsari, A., & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*, 121-127. https://doi.org/10.1136/amiajnl-2011-000089 (2012).
13. Romeo, G., & Conti, D. Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. *AI & Soc*; 10.1007/s00146-025-02422-7 (2025).
14. Dror, I. E. Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. *Analytical Chemistry*, **92,** 7998-8004; 10.1021/acs.analchem.0c00704 (2020).
15. Logg, J., Minson, J. & Moore, D. Algorithmic appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* **151,** 90–103 (2019).
16. Kidd, C., & Birhane, A. How AI can distort human beliefs. *Science.* **380,** 1222-1223; 10.1126/science.adi0248 (2023).

17. Leffer, L. *Humans Absorb Bias from AI – And Keep It after They Stop Using the Algorithm.* Scientific American. https://www.scientificamerican.com/article/humans-absorb-bias-from-ai-and-keep-it-after-they-stop-using-the-algorithm/ (2023).
18. van der Miesen, M. M., van der Lande, G. J. M., Hoogeveen, S., Schjoedt, U., & van Elk, M. The effect of source credibility on the evaluation of statements in a spiritual and scientific context: A registered report study. *Comprehensive Results in Social Psychology*, *6*(1–3), 59–84; 10.1080/23743603.2022.2041984 (2022).
19. Sabbagh, M. A., & Baldwin, D. A. Learning Words from Knowledgeable versus Ignorant Speakers: Links Between Preschoolers' Theory of Mind and Semantic Development. *Child Development*, **72,** 1054-1070; 10.1111/1467-8624.00334 (2003).
20. Placani, A. Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, **4**, 691-698; 10.1007/s43681-024-00419-4 (2024)
21. Birhane, A., & van Dijk, J. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Association for the Advancement of Artificial Intelligence), 207–213 (2020)
22. Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., & Agreda, S. G. To trust or not to trust? AN assessment of trust in AI-bases systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change*, **181**; 10.1016/j.techfore.2022.121763 (2022).
23. Sanders, T., Kaplan, A., Koch, R., Schwarts, M., & Hancock, P. A. The Relationship Between Trust and Use Choice in Human-Robot Interaction. *Human Factors*, **61**(4), 614-626; 10.1177/0018720818816838 (2019).
24. Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, **58**(3), 377-400; 10.1177/0018720816634228 (2016)
25. Nickerson, J. V., & Reilly, R. R. A model for investigating the effects of machine autonomy on human behaviour. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*; 10.1109/HICSS.2004.1265325 (2004).
26. Konnikova, M. The Hazards of Going on Autopilot. *The New Yorker.* https://www.newyorker.com/science/mario-Konnikova/hazards-automation (2014).
27. Dror, I. E., Wertheim, K., Fraser-Mackenzie, P., & Walajtys, J. The impact of human-technology cooperation and distributed cognition in forensic science: Biasing effects of AFIS contextual information on human experts. *Journal of Forensic Sciences*, **57,** 343-352 (2012).
28. Dror, I. E., & Mnookin, J. The use of technology in human expert domains: Challenges and risks arising from the use of automated fingerprint identification systems in forensics. *Law, Probability and Risk*, **9,** 47-67 (2010).
29. Nightingale, S. J. & Farid, H. AI-Synthesized Faces are Indistinguishable from Real Faces and More Trustworthy. *Proceedings of the National Academy of Sciences*, **119**; 10.1073/pnas.2120481119 (2022).
30. Karras, T., Laine, S., & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *Computer Science: Neural and Evolutionary Computing*. (2019).
31. Frazier, M. L., Johnson, P. D., & Fainshmidt, S. Development and validation of a propensity to trust scale. *Journal of Trust Research*, **3,** 76-97; 10.1080/21515581.2013.820026 (2013).

32. Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. Measuring trust. *Quarterly Journal of Economics* **115,** 811–46; 10.1162/003355300554926 (2000).
33. Rotter, J. B. A new scale for the measurement of interpersonal trust. *Journal of Personality*, **35,** 651–665; 10.1111/j.1467-6494.1967.tb01454.x (1967).
34. Yamagishi, T. The provisioning of a sanctioning system as a public good. *Journal of Personality and Social Psychology,* **51,** 110-116; 10.1037/0022-3514.51.1.110 (1986).
35. Yamagishi, T. The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly* **51,** 265–71 ; 10.2307/2786924 (1988).
36. Yamagishi, T. & Yamagishi, M. Trust and commitment in the United States and Japan. *Motivation and Emotion,* **18,** 129-166; 10.1007/BF02249397 (1994).
37. Nunnally, J. C.,& Bernstein, I. H. Psychometric theory (3rd ed.) (McGraw-Hill, 1994).
38. Schepman, A., & Rodway, P. Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, **1,** 100014; 10.1016/j.chbr.2020.100014 (2020).
39. Green, D. M., & Swets, J. A. *Signal detection theory and psychophysics* Vol. 1, 1969-2012 (New York: Wiley, 1966).
40. Macmillan, N. A., & Creelman, C. D. *Detection theory: A user's guide* (Psychology press 2004) 10.4324/9781410611147
41. Cohen, J. *Statistical power analysis for the behavioral sciences* (2nd ed.), (Hillsdale, NJ: Lawrence Erlbaum, 1988).
42. Makowski, D. *The psycho Package: an Efficient and Publishing-Oriented Workflow for Psychological Science*. Journal of Open Source Software, **3,** 470; 10.21105/joss.00470 (2018).
43. Lerman, D. C. et al. Applying signal-detection theory to the study of observer accuracy and bias in behavioural assessment. *Journal of Applied Behaviour Analysis*, **43**; 195-213; 10.1901/jaba.2010.43-195 (2010).
44. Hu, L. T., & Bentler, P. M. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling*, **6**, 1-55; 10.1080/10705519909540118 (1999).
45. Alavi, M., Visentin, D. C., Thapa, D. K., Hunt, G. E., Watson, R., & Cleary, M. Chi-square for model fit in confirmatory factor analysis. Journal of advanced nursing, 76(9), 2209-2211. https://doi.org/10.1111/jan.14399 (2020).
46. Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.) (Guilford Press, 2005).
47. West, S. G., Taylor, A. B., & Wu, W. Model fit and model selection in structural equation modelling in *Handbook of structural equation modeling* (ed. Hoyle, R. H.) 209-231 (Guilford Press, 2012).
48. Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, **21,** 422–430; 10.1016/j.concog.2011.09.021 (2012).
49. Maniscalco, B., & Lau, H. Signal detection theory analysis of type 1 and type 2 data: meta-d', response-specific meta-d', and the unequal variance SDT mode in *The Cognitive Neuroscience of Metacognition* (eds. Fleming, M., & Frith, C. D) 25-66 (Springer, 2014).
50. Schoeffer, J., De-Arteaga, M., & Kühl, N. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the CHI Conference on*

Human Factors in Computing Systems (CHI '24), 1-18; 10.1145/3613904.3642621 (2024).
51. Wickens, C. D., Helton. W. S., Hollands, J. G., & Banbury, S. Chapter 13 | human-automation interaction in *Engineering psychology and human performance* 516-551 (Routledge, 2021).
52. Carragher, D. J., & Hancock, P. J. B. Simulated automated facial recognition systems as decision-aids in forensic face matching tasks. *Journal of Experimental Psychology: General*, **152,** 1286–1304; 10.1037/xge0001310 (2023).
53. Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. Explaining models: an empirical study of how explanations impact fairness judgement. In *Proceedings of the 24th international conference on intelligent user interfaces* 275-285 (2019).
54. Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. Does Explainable Artificial Intelligence Improve Human Decision-Making? In *Proceedings of the AAAI Conference on Artificial Intelligence,* **35,** 6618-6626; 10.1609/aaai.v35i8.16819 (2021).
55. Zhang, Y., Liao, Q. V., & Bellamy, R. K.. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency,* 295-305 (2020).
56. Gino, F., & Moore, D. A. Effects of task difficulty on use of advice. *Journal of Behavioural Decision-Making*, **20**, 21-35; 10.1002/bdm.539 (2007).
57. Bogert, E., Schechter, A., & Watson, R. T. Humans rely more on algorithms than social influence as a task becomes more difficult. *Sci Rep,* **11**, 8028; 10.1038/s41598-021-87480-9 (2021).
58. Madhavan, P., Wiegmann, D. A. and Lacson, F. C. Automation failures on tasks easily performed by operators undermines trust in automated aids. *Human Factors*, **48,** 241-256 (2006).
59. Carragher, D. J., Sturman, D., & Hancock, P. J. B. Trust in automation and the accuracy of human-algorithm teams performing one-to-one face matching tasks. *Cognitive Research: Principles and Implications,* **9,** 41; 10.1186/s41235-024-00564-8 (2024).
60. Palermo, R. et al. Do people have insight into their face recognition abilities? *The Quarterly Journal of Experimental Psychology,* **70,** 218-233; 10.1080/17470218.2016.1161058 (2017).
61. Flowe, H. D. et al. An experimental examination of the effects of alcohol consumption and exposure to misleading postevent information on remembering a hypothetical rape scenario. *Applied Cognitive Psychology,* **33,** 393–413; 10.1002/acp.3531 (2019).
62. Kruger, J., & Dunning, D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, **77,** 1121–1134; 10.1037/0022-3514.77.6.1121 (1999).

Author contributions statement

E. J. and S. N. conceived the experiment. E.J., I. D., G. R. W., G. M., and S. N. designed the experiment. J. P. conducted the experiment, analysed results, and wrote the manuscript. All authors reviewed the manuscript.

Additional information

Competing interests statement. The corresponding author confirms that there are no competing interests to declare.

Funding. This research was funded by the UK Defence Science and Technology Laboratory (Dstl) through The Alan Turing Institute's AI Research Centre for Defence (ARC-D). All views expressed in this report are those of the authors, and do not necessarily represent the views of Lancaster University, The Alan Turing Institute or any other organisation.