

Trusting an Algorithm Can Be a Tricky and Sticky Thing

Garston Liang^{1, 2}, Amy X. Li^{1, 3}, and Ben R. Newell¹

¹ School of Psychology, University of New South Wales Sydney

² School of Psychological Sciences, University of Newcastle

³ Department of Experimental Psychology, University of Oxford

What information guides individuals to trust an algorithm? We examine this question across four experiments that consistently found *explanations* and *relative performance* information increased ratings of trust in an algorithm relative to a human expert. When participants learn of the algorithm's shortcomings, we find that trust can be broken but, importantly, restored. Strikingly, despite these increases and restorations of trust, few individuals changed their overall preferred agent for each commonplace task (e.g., driving a car), suggesting a conceptual ceiling to the extent to which people will trust algorithmic decision aids. Thus, initial preferences for an algorithm were “sticky” and largely resistant, despite large numeric shifts in trust ratings. We discuss theoretical and practical implications of this work for researching trust in algorithms and identify important contributions to understanding when information can improve people's willingness to trust decision aid algorithms.

Keywords: algorithm, accuracy, expert

Supplemental materials: <https://doi.org/10.1037/dec0000229.supp>

Machine learning algorithms touch nearly all aspects of daily life. Humming in the background, algorithms are trained on vast quantities of data to filter spam from our inboxes, detect suspicious payment transactions, and translate speech to text in real time. Beyond the mundane, algorithms also assist doctors in medical imaging, assess recidivism risks for bail sentencing, and pilot autonomous vehicles. Seemingly, if the data exist, clever

analysts can craft machine learning algorithms that make predictions about the future. However, whether these predictions are recruited or ignored depends on convincing decision-makers of the merits of algorithmic recommendations.

In this article, we investigate what information convinces individuals to trust an algorithm. Across four experiments, we present a series of scenarios with familiar tasks, such as driving a vehicle, in which participants indicate their preference for either an algorithm or human agent to perform each task. After making an initial rating, we then introduce summaries from published academic studies where machine learning algorithms outperformed their human counterparts. With this new information in mind, we sought to understand whether individuals revise their trust in the algorithm. To preface the results, we show that information about (a) performance evaluations of the algorithm and (b) explanations of how the algorithm operates increases people's numeric ratings of trust in the algorithm. When we present individuals with both pieces of information, however, we find that increases in trust hit a conceptual ceiling.

This article was published Online First March 21, 2024.

Garston Liang  <https://orcid.org/0000-0002-9230-7258>

The authors thank Associate Professor Castelo for kindly sharing the original data from Castelo et al. (2019). The support of the Australian Research Council (DP190101076; DP190101675) is gratefully acknowledged. All analysis code and data are available at this Open Science Framework archive: <https://osf.io/dsfg9/>.

Garston Liang played a lead role in conceptualization, data curation, formal analysis, and writing—original draft and an equal role in writing—review and editing. Amy X. Li played a supporting role in data curation, formal analysis, and writing—review and editing. Ben R. Newell played a lead role in conceptualization, supervision, and writing—review and editing.

Correspondence concerning this article should be addressed to Garston Liang, School of Psychology, University of New South Wales Sydney, Kensington, Sydney, NSW 2052, Australia. Email: garston.liang@gmail.com

Complementing this ceiling, we find these initial preferences are sticky—only a fraction of individuals changed their overall preferred agent despite large quantitative changes in trust. Together, these results shed light on the extent to which algorithm information encourages individuals to recruit algorithmic recommendations into their decision making.

Critical to whether decision-makers use algorithms is what information they possess about the algorithm's capabilities. This information may be learnt through direct experience, such as test-driving the autopilot feature on a Tesla, or acquired through descriptions of the algorithm, such as safety reports that validate the system's reliability. We focus on the latter descriptive form of information about the algorithm and how this description affects the trust individuals place in its capacities.

Experiments in both applied and experimental settings have observed that individuals often reject the advice of algorithms. Collectively, these observations have been termed "algorithm aversion" when individuals with direct experience of an algorithm's imperfect recommendations opt against seeking its advice, even when that algorithm outperforms the individual's own capabilities (for review, see [Burton et al., 2020](#); [Dietvorst et al., 2015](#)). Recent attention has focused on how different information about the algorithm affects this degree of aversion ([Liang et al., 2022](#)). For example, when individuals were given positive descriptions of the algorithm's capabilities prior to any direct experience, they held favorable attitudes toward using its recommendations ([Logg et al., 2019](#)).

Performance Information

To help a person decide whether to use an algorithm, a key piece of information might be how the algorithm's performance compares to that of a human operator. Along this vein, [Castelo et al. \(2019\)](#) created nine summaries that directly compared algorithms to human agents. Each summary provided empirical evidence of different scenarios in which algorithms outperformed human agents, ranging from movie recommendations to medical treatment planning. For example, autonomous vehicles were found to be 28% safer than human drivers ([Blanco et al., 2016](#)), while a machine learning system predicted ratings for jokes with 7% greater accuracy

([Yeomans et al., 2019](#)). In an experiment, [Castelo et al. \(2019\)](#) presented these nine summaries to participants, who then indicated their trust for either a human or algorithmic agent to perform each individual task. In a between-subjects design, Castelo et al. found that providing performance information increased ratings of trust in the algorithm compared to when the information was absent.

An open question is whether the degree of algorithm superiority (compared to humans) affects people's trust in algorithmic recommendations. Presumably, studies that showed larger improvements with an algorithm provide the grounds for greater trust in its capabilities. As the degree of algorithm superiority and scenario content were not independently manipulated, Castelo et al. could only indirectly assess this prediction across qualitatively different scenarios, e.g., comparing the 7% algorithm improvement for joke-telling to the 28% improvement for driving safety. In our experiments, we directly manipulate the improvement information within the same scenarios to examine the impact of performance information on people's trust in algorithms.

Explanations of the Algorithm's Process

When seeking advice from another person, the individual can walk through their reasoning to explain their perspective (provided you asked nicely). Communicating the specific steps that generated a particular algorithmic recommendation, however, is very likely more involved. Between the large training data sets, hidden layers in the model's architecture, and the underlying statistical methods, it is possible that some of the reluctance people experience with algorithms arises from a lack of understanding about how these recommender systems work.

In our experiments, we examined how providing an explanation of the algorithm affects ratings of trust. Even if an algorithm was superior to human performance, decision-makers may still want an explanation if they are concerned about safety trade-offs, the ethics of a decision process, or the alignment with the user's actual needs and the algorithm's optimization function ([Doshi-Velez & Kim, 2017](#)). These broad concerns have sprouted new subfields of research in Explainable Artificial Intelligence and prompted wide-reaching protection laws such as the European Union's

General Data Protection Regulations (Shin, 2021). Our experiments explore this core idea that trust is founded upon transparency of process.

We directly compare simple explanations to more complex versions to understand how the level of explanatory detail may play a role. Previous work by Dzindolet et al. (2003) found that even simple explanations about why an algorithmic aid could fail improved people's trust and use of the algorithm in a subsequent image detection task. However, simple explanations may backfire if individuals hold prior expectations that the algorithm should be more complex (Lehmann et al., 2022). Indeed, much resistance within clinical circles to the statistical decision rules advocated by Meehl (1954) was centered on the fact that their simplicity belied the complexities of diagnosis.

More recently, there is evidence that people prefer recommender systems that provide additional reasons to support a recommendation (Rago et al., 2021). However, trust can be compromised should these reasons place too high a burden on the decision-maker to mentally simulate a complex algorithm (Poursabzi-Sangdeh et al., 2021). We tested this possibility by presenting both simple and more complex explanations of the algorithm's process.

In our experiments, we employ the Judge-Advisor-System methodology that first obtains a baseline rating of trust before observing how new information affects the decision-maker's preferences when they update their baseline ratings. One benefit of this methodology is that we can determine whether explanations of any kind affect ratings of trust at all. Indeed, across different contexts, a number of studies have found explanations add little benefit to behavioral outcomes with algorithms (Lehmann et al., 2022; Schmidt et al., 2020).

Experiment Overview

We adapted the between-subjects experiment from Castelo et al. (2019) to present algorithm information in a within-subjects design. Castelo et al. found that individuals who received algorithm information had higher ratings of trust compared to groups where this information was absent. Our within-subject extension seeks to expand the purview of algorithm information in four ways to examine: (a) the degree of

algorithm superiority over the human expert in Experiment 1, for example, 20% better versus 40% better; (b) the complexity of explanations about how the algorithm operates in Experiment 2, that is, simple 1-step explanations versus multistep explanations; (c) the impact and order of information in combination in Experiment 3, that is, superiority information first versus explanation information first; and (d) information order when learning about algorithm trade-offs in Experiment 4, that is, benefits versus pitfalls of algorithmic advice.

We hypothesized that (a) a larger degree of performance superiority in favor of the algorithm would lead to larger increases in trust ratings (Experiment 1) (b) that complex explanations would lead to greater increases in trust than simple explanations (Experiment 2), and (c) that presenting both pieces of information (i.e., performance superiority and explanations) would lead to a subadditive change in how willing participants were to trust an algorithm. The added benefit of our within-subjects design allows us to examine how each parcel of information affects individuals at the subject-level.

General Method

We first describe general details that were common across all four experiments before proceeding to experiment-specific Methods and Results.

Participants

Experiments 1, 2, and 3 recruited 150 participants for each experiment (per condition $n = 75$, $N = 450$; $N_{\text{male}} = 136$, $N_{\text{female}} = 299$, $N_{\text{other gender}} = 12$, $N_{\text{prefer not to say}} = 1$; $N_{\text{failed to record}} = 2$; $M_{\text{age}} = 27.9$), and Experiment 4 recruited 602 participants ($n \geq 100$, $N_{\text{female}} = 254$, $N_{\text{male}} = 338$, $N_{\text{nonbinary}} = 6$, $N_{\text{other}} = 4$; $M_{\text{age}} = 38.72$). Our sample size was chosen to replicate and extend Castelo et al. (2019) original between-subjects design, recruiting a similar number of participants with additional statistical power from a within-subjects design.

Participants were recruited from Prolific Academic with three exclusion criteria: that participants exceeded an 80% approval rating, that English was their primary language, and that they had not participated in our previous experiments of the same design. Participants were remunerated £1.66 for participation in the

10-min experiment, and the ethics protocol was approved by the University of New South Wales Sydney Human Research Ethics Advisory Panel.

General Materials

The experiment was programmed in Javascript using JSPsych (De Leeuw, 2015) and completed on a webpage.

Trust Ratings

Participants inputted their ratings of trust on a slider scale to the question “Who would you trust more to do [X]?” where X is a given scenario described in the materials subsection. The slider scale spanned from 0 to 100 with labels for algorithm (0), equal trust in algorithm/human (50), and human expert (100; see Panel A of Figure 1). By default, the slider start point was set to the midpoint of the scale at 50. Participants could proceed only once they had clicked on the slider scale. For subsequent ratings, the start point was updated to the individual’s previous trust rating and did not require any interaction to proceed if they wished to retain that rating.

Scenarios and Algorithm Information

Our experiments presented four scenarios where computational algorithms have been validated against human performance in past studies. The four scenarios were as follows: (a) recommending a medical treatment plan (Lohr, 2016), (b) safety when driving a car (Blanco et al., 2016), (c) funniness of a joke (Yeomans et al., 2019), and (d) recommending a movie (Krishnan et al., 2008). We selected these four scenarios from Castelo et al. (2019). Although tangential to our pursuits, we used Castelo et al.’s objectivity ratings¹ to guide our selection of the two highest (medical, driving scenarios) and two lowest objectivity-rated scenarios (joke-telling, movie recommendations). Our intent was to identify a range of situations where initial preferences for algorithmic decision making may be more favorable (self-driving vehicles and movie recommendations) compared to areas traditionally reserved for human judgment (joke-telling and medical diagnosis).

For each scenario, we provided text summaries of the above-cited empirical validation studies. This text summarized key information about the

algorithm; we present these summaries in full in the Appendix. For expositional purposes, we provide an abridged version of the automated driving scenario and the associated text summaries in Table 1. Bookending the algorithm information were standardized prompts to elicit initial and subsequent ratings of trust along the slider scale. Across the scenarios, the prompts and performance validation information were identical to those in Castelo et al. (2019) that demonstrated text summaries increased trust in an algorithm, notably using a between-subject design. Our within-subject extension retains all aspects of the previous study, with the exception that webpage links that were originally embedded into the text summaries could not be provided because we directly manipulated the content of the summaries.

Experiment 1, 2, and 3

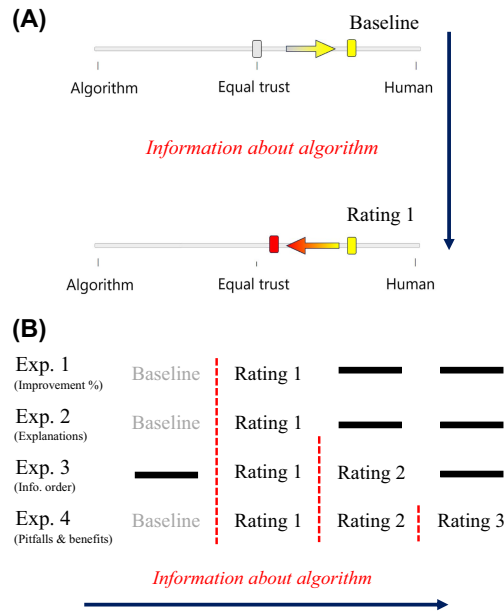
Design

Experiments 1, 2, and 3 used a two-group between-by-within-subject design. Scenario and rating instance were within-subject factors where the presentation order of the four scenarios was randomized. The between-subject factor differed by experiment and is illustrated in Panel B of Figure 1. Experiment 1 manipulated the stated degree of improvement of the algorithm, presenting either the veridical improvement percentage or doubled improvement (e.g., the 28% or 56% safer in the autonomous driving scenario). Experiment 2 manipulated the explanation complexity related to how the algorithm generated its recommendation (simple vs. complex²). Experiment 3 combined the veridical improvement percentage from Experiment 1 and the complex explanations from Experiment 2 and manipulated the order

¹ Castelo et al. (2019) asked participants to rate a given task, for example, providing movie recommendations, as objective/subjective on a scale ranging from 1 to 100, based on the description of “objectivity” as how quantifiable and measurable a successful outcome might be and “subjectivity” defined as how open to interpretation and opinion a given task may be.

² Separate pilot ratings ($N = 100$) confirmed that subjects overwhelmingly perceived the difference in explanation complexity. Complexity of explanation statements were ranked from 1 (*simplest*) to 4 (*most complex*) for each scenario (mean rank for simple vs. complex; medical: 1.5 versus 3.7, driving: 1.2 versus 3.8, joke: 1.4 versus 3.9, movie: 1.3 versus 3.8; all $ps < .001$, nonparametric Wilcoxon signed-rank test).

Figure 1
Task Description and Experiment Outline



Note. Panel A shows an example of a scenario and how participants inputted and updated their ratings of trust. The downward black line indicates the progression of the scenario. Initially, the response slider defaults to the midpoint and is moved to the yellow slider position, indicated by the yellow-colored arrow. After this baseline rating is entered, algorithm information was presented, and participants were prompted to update their trust for Rating 1. The slider is shown again and begins at the same point as the previous rating, that is, the yellow slider position. In the example, the participant shifts their rating back toward the “equal trust” position following the presentation of the information. This updating procedure is repeated when more than one piece of algorithm information is presented (i.e., Experiments 3 and 4). Panel B provides an overview of the experiment flow. The text in each cell shows when participants made a trust rating for each experiment. Red vertical lines between ratings show the point at which information about the algorithm was displayed to the participant. Black dashes indicate absent ratings, and the long arrow underneath indicates scenario progression. Exp. = experiment. See the online article for the color version of this figure.

of information presentation between-subjects (accuracy-first, explanation-first).

Procedure

Participants were instructed that their task was to indicate their trust in either an algorithm or a human judge in four different scenarios. Participants first provided an initial numerical trust rating to the prompt “Who would you trust

more [in scenario X]?” before being presented with algorithm information in the form of a study summary. The conditions diverged based on the content of these study summaries (for content, see Table 1 and Appendix; for conditions, see Panel B of Figure 1).

After reading the algorithm information, participants were prompted to update their ratings of trust, where the slider start position was set to their previous ratings. In Experiment 1 and Experiment 2, the *baseline rating* was made in the absence of any study summary. In Experiment 3, the first rating was made with algorithm information present (see Panel B of Figure 1). As such, each subject completed the experiment after providing two ratings for each of the four scenarios.

Results Experiment 1, 2, and 3

Analysis was conducted in R (R Core Team, 2015) using the lme4 package (Bates et al., 2015). We fit linear mixed models with information condition and scenario as fixed factors and random slopes for subjects. We calculated difference scores from both trust ratings (i.e., initial rating trust—postrating = Δ trust scores) and reported the means and standard errors with each analysis. Follow-up tests use Bonferroni-corrected p values.

We present the results of Experiments 1, 2, and 3 together in Figure 2. Panel A consists of three colored rows, one for each experiment and two columns for each manipulation (Experiment 1: actual % vs. doubled %, Experiment 2: simple vs. complex explanation, Experiment 3: order of information %-first or explanation-first). Within each of these between-subjects condition, we present all four scenarios along the x -axis, with symbols indicating matching scenarios. We begin by describing the algorithm information effect, that is, trust increased following algorithm information summaries, in order of each experiment, beginning with the orange-colored row for Experiment 1.

Algorithm Information Effect (Experiment 1)

In Experiment 1, trust in the algorithm increased following algorithm performance information about the algorithm’s improvement over a human operator, relative to the preinformation baseline (orange upper panels in Figure 2, Panel A; H_1 : Δ trust scores $\neq 0$ vs. H_0 : Δ trust

Table 1

Algorithm Information and Response Prompts for the Automated Driving Scenario Used in All Four Experiments

Information	Content
Rating prompt	Who would you trust more to drive a car safely?
Information preamble	A recent study conducted by professional academic researchers showed that cars driven by algorithms experience fewer accidents compared to human drivers.
Veridical versus doubled performance improvement	In the study, cars driven by algorithms were 28% (or, 56%) safer than cars driven by humans.
Simple explanation	When the algorithm detects danger, it can brake faster than a human driver can.
Complex explanation	The algorithm used a system of cameras to predict if an object in the environment would collide with the car. If there was the possibility of danger, the algorithm calculated the object's distance and took precautionary action (e.g., braking or changing lanes) faster than a human driver could.

Note. The veridical performance improvement (in bold), prompt, and preamble were matched for wording to [Castelo et al. \(2019\)](#).

scores = 0, all $ps < .004$). Notably, we did not find evidence that doubling the performance of the algorithm affected trust ratings (left-orange vs. right-orange panels, $M_{\Delta \text{ trust, doubled-\% condition}} = 12.8$ vs. $M_{\Delta \text{ trust, actual-\% condition}} = 15.4$; $\chi^2(1) = 1.90$, $p = .17$). Trust in the algorithm increased the most in the medical diagnosis scenario ($M_{\Delta \text{ trust}} = 20.5$, $SE = 2.16$) compared to all other scenarios ($\chi^2(3) = 34.36$, $p < .001$). We find evidence of a scenario by accuracy interaction (left vs. right orange panels, $\chi^2(3) = 21.73$, $p < .001$). This interaction was likely driven by lower trust ratings in the movie scenario in the actual-% condition (left-orange panel, $\Delta \text{ trust}$ in movie scenario vs. other scenarios, all $ps < .005$) which did not differ from other scenarios in the doubled-% condition (right-orange panel, all $ps > .14$; $\chi^2(3) = 9.31$).

Algorithm Information Effect (Experiment 2)

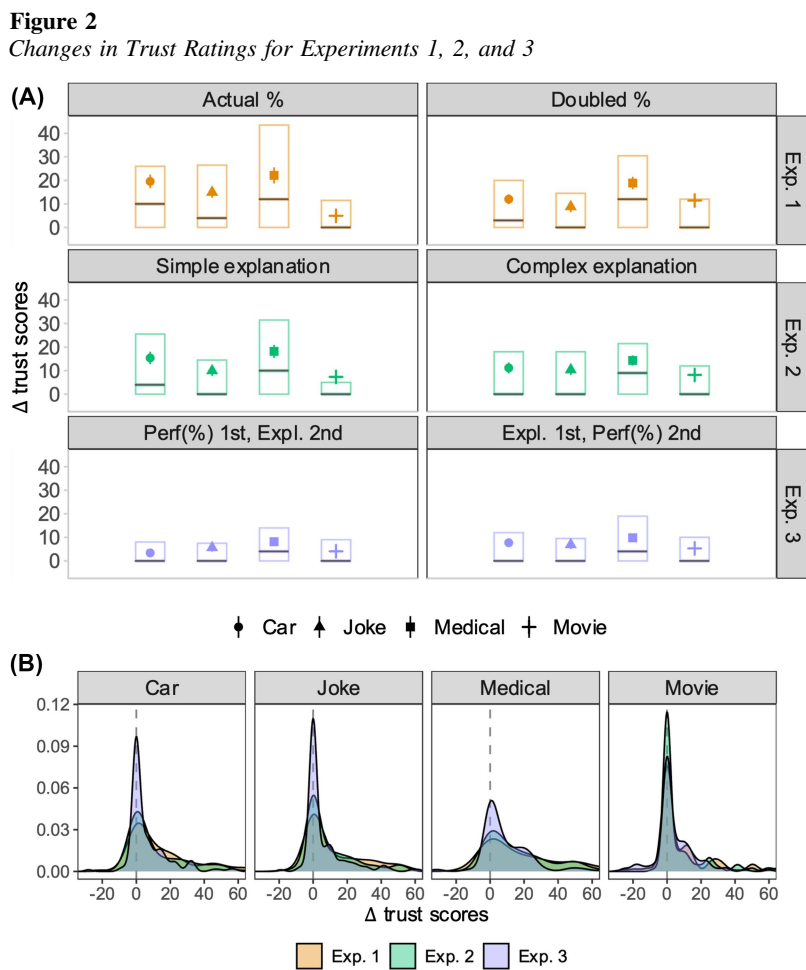
In Experiment 2, trust in the algorithm increased following explanations about how the algorithm operates (green panels in [Figure 2](#); H_1 : $\Delta \text{ trust scores} \neq 0$ vs. H_0 : $\Delta \text{ trust scores} = 0$, all $ps < .003$). Contrary to our expectations, we did not find evidence that trust increased more for complex explanations compared to simple explanations (green-left panel vs. green-right panel, $M_{\Delta \text{ trust, complex condition}} = 11.0$, $SE = 1.07$ vs. $M_{\Delta \text{ trust, simple condition}} = 12.7$, $SE = 1.23$; $\chi^2(1) = 0.85$; $p = .36$). As in Experiment 1, the scenario that prompted the greatest increase in trust for the algorithm was the medical scenario ($M_{\Delta \text{ trust}} = 16.3$, $SE = 1.75$, $t(149) = 9.30$, $p < .001$), with the smallest increase in the movie recommendation

scenario ($M_{\Delta \text{ trust}} = 7.75$, $SE = 1.35$, $t(149) = 3.94$, $p < .001$). A tentative explanation for this pattern is that trust was initially higher for the movie recommendation scenario (see [Supplemental Material](#) for initial ratings). It is possible that the everyday familiarity with media recommendation algorithms meant that adding explanation information did not shift the existing high level of trust.

Algorithm Information Effect (Experiment 3)

Overall, Experiments 1 and 2 showed that both statements containing performance information and explanations can increase trust in algorithms. In Experiment 3, we manipulated the order of presentation for each statement. We expected the improvement-first condition may have smaller increases in trust relative to the explanation-first condition because, after one learns the algorithm is superior to the improvement information, subsequently learning the explanation may only be a curiosity.

In Experiment 3 (purple third row, [Figure 2](#)), cumulative algorithm information (i.e., containing both performance information and explanations) increased trust in the algorithm across three scenarios (all $ps < .04$), with the exception of the improvement-first, movie recommendation scenario (cross-symbolled column, $M_{\Delta \text{ trust}} = 4.07$, $SE = 2.09$; $p = .06$). At least numerically, the greatest increase in trust was in the medical scenario ($M_{\Delta \text{ trust}} = 8.97$, $SE = 1.30$, $p < .001$) and lowest in the movie recommendation scenario ($M_{\Delta \text{ trust}} = 4.69$, $SE = 1.31$, $p = .002$). We failed to observe differences between the information



Note. Panel A shows changes in trust from baseline trust ratings following algorithm information for all six conditions. Positive values denote increases in trust in an algorithm, and conversely, decreasing scores represent decreasing trust in the algorithm in favor of a human operator. Each scenario is presented separately as symbols, where each scenario is plotted with means and standard errors, boxplots for the central interquartile range, a black median line within the boxplot, and color showing the experiment number. Panel B presents a density plot of individual-level trust-change scores for each scenario separated by panels, and the experiment is shown in color. Vertical dotted intercept line at $x = 0$ shows 0-change in trust between baseline ratings of trust and the ratings after the algorithm information. Units on the y-axis show density and units on the x-axis are commensurate with aggregate-level data plotted in Panel A. Exp. = experiment; Expl. = explanation; Perf. = performance. See the online article for the color version of this figure.

order conditions (left-purple panel vs. right-purple, $\chi^2(1) = 2.47, p = .12$), scenarios (within panels, $\chi^2(3) = 7.02, p = .07$), or an interaction ($\chi^2(3) = 1.14, p = .77$).

Compared with previous experiments, the mean change in trust ratings in Experiment 3 was numerically smaller. To understand better the cross-experiment comparison, we plot the

distribution of trust-change scores in Panel B of Figure 2. One observation from the distributional data is that Experiment 3 recorded more zero-change scores across each of the scenarios compared to Experiments 1 or 2, indicating that many participants retained their initial trust rating after viewing the algorithm information (Panel B of Figure 2; purple distribution peaked

at 0). This 0-peak likely reflects the fact that additional information in the initial rating raised the baseline level of trust, creating a ceiling effect that reduced the space for movement for the second piece of algorithm information (see [Supplemental Figure S1](#) for initial trust scores).³

Interim Discussion

Our first three experiments showed that providing performance information and explanations, in combination and independently, increased people's trust in an algorithm. We observed these increases across four different scenarios in which we presented information from academic studies where the algorithm outperformed a human operator. These results extend the algorithm information effect reported by [Castelo et al. \(2019\)](#) to help elucidate the elements of algorithm-related information that convince decision-makers to trust in an algorithm. Interestingly, however, the positive effect of algorithm information on trust was not additive. The results of Experiment 3 hinted at a conceptual ceiling to which trust scores could change in response to more information about the same algorithm.

In pursuit of understanding the limits of algorithm information, we designed Experiment 4 with two changes in mind. The first was to extend the number of ratings and information summaries within a single scenario, namely the automated driving scenario. Examining the effect of information in closer detail within a single scenario allowed us to expand our previously brief descriptions of the autonomous driving algorithm while also examining the limits of information summaries in response to three ratings rather than one (see [Figure 1](#) for diagram).

The second change was to introduce *pitfall information*, which describes possible algorithm limitations. Across earlier experiments, the information participants received about algorithms was always in favor of the algorithm; however, there are a number of real-world examples where trusting in an algorithm may incur substantial costs, ranging from systematic bias in hiring decisions ([Dastin, 2018](#)) to discriminatory bail terms based on selection effects in training data ([Angwin & Larson, 2016](#)). To incorporate these real-world considerations, we introduce negative information describing where algorithm-based systems might fail when a human would not. In addition to eliciting a higher number of trust

ratings and providing more in-depth algorithm information, Experiment 4 therefore allows us to examine how people's trust in algorithms responds to a more balanced view of both an algorithm's benefits and pitfalls.

One open question is how negative information is interpreted if individuals learnt of the algorithm's benefits beforehand. It is possible that learning about the relative benefits of a new system makes subsequent information about its costs more palatable. Indeed, the reverse case, where the negative costs are presented first, may persuade an individual that the new technology is inherently flawed. Separately, the pitfall information also provides a context for interpreting any performance comparison (e.g., information that an algorithm is 30% safer than a human agent). An individual who learns that an algorithmic system is still safer than a human driver, despite the potential for infrequent errors, might be more trusting of the algorithm compared to if they had learnt about the performance comparison on its own.

Therefore, in Experiment 4, we manipulate the order in which participants learn information about the benefits versus pitfalls of an algorithm. Importantly, the presentation order where the pitfall information is encountered prior to the benefit-related information allows us to examine whether trust that is lost can be restored. By manipulating these different trajectories of algorithm-related content, we can examine the different paths that individuals traverse when deciding to trust an algorithm.

Experiment 4

Design

Participants provided an initial rating regarding their trust in autonomous vehicles. They were then presented with three pieces of information and updated their trust rating after seeing each piece (see [Figure 1B](#)). The primary between-subjects manipulation was that of information

³ One explanation of the larger proportion of 0-change scores is that participants were less engaged with task and defaulted to retaining their initial trust ratings. We rule out this explanation in the General Discussion (see [Supplemental Figure S4](#)) and show that a consistent majority (~85% of participants in each experiment) changed ratings to at least one scenario.

order—for example, pitfalls then benefits then performance comparison versus benefits then performance comparison then pitfalls—yielding a total of six unique information-order conditions. Allocation to information-order condition was randomized between participants (n per condition = 100).

Materials

Single Scenario

Experiment 4 was restricted to the automated driving scenario used in earlier experiments. This scenario was chosen because we expected our participants to be relatively familiar with the potential benefits and risks of automated driving.

Algorithm Information

Participants saw three pieces of information: (a) the performance comparison, (b) benefits of automation, and (c) potential pitfalls. The performance comparison statement was identical to that used in previous experiments. The benefits information detailed how algorithmic systems in automated driving had eliminated common human-error accidents such as mistaking pedals while parking (e.g., depressing the accelerator instead of the brake). The potential pitfalls information detailed how algorithmic-based systems may fail to recognize abnormal driving scenarios where human intervention would avoid fatal consequences.

Procedure

Participants were informed that their task was to report their trust in algorithm-based driving systems. They were presented with an initial statement about a recent academic study into automated vehicles and how algorithms contributed to their functioning. Participants made an initial rating of trust on the scale.

Following the initial rating, all participants received three pieces of information about the benefits of the algorithm, the pitfalls, and the performance comparison, the order of which was randomized according to experimental condition. Once one rating was submitted, the new piece of algorithm information was then appended to the same page and highlighted such that all previous information about the algorithm remained visible

on the page. Participants completed the experiment after inputting their four ratings.⁴

Results—Experiment 4

We present the results for Experiment 4 in two parts. In the first section, we analyze the numeric trust scores and show that information content (Figure 3, Panel A) and order (Figure 3, Panel B) affected trust in an autonomous driving algorithm. In the second section, we use the coarser binary preference metric to show that despite these numeric increases in reported trust, the proportion of individuals that changed their overall preferred agent was small.

As an initial manipulation check of the pitfall information, trust changed as a function of the information content (Panel A of Figure 3, $\chi^2(10) = 35.91$; $p < .001$). Across information-order conditions, trust ratings increased after viewing the benefits ($M_{\Delta \text{ trust}} = 6.04$, $SE = 0.52$) and performance information ($M_{\Delta \text{ trust}} = 8.03$, $SE = 0.65$) and decreased after viewing the pitfall information ($M_{\Delta \text{ trust}} = -14.70$, $SE = 0.74$, all Bonferroni-adjusted $ps < .001$).

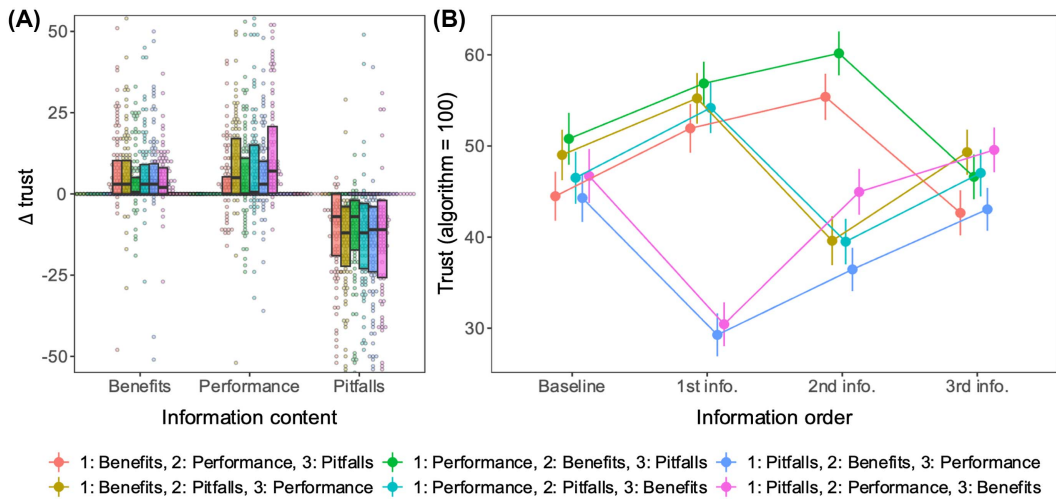
Because the order of the information interacted with information content ($\chi^2(10) = 35.91$; $p < .001$), we examine the trust ratings over the time course of the experiment in Panel B of Figure 3 and extract three core findings. The first is that, at baseline, only a single group favored the autonomous driving algorithm over a human driver, albeit marginally (Figure 3 green-color order, $M_{\text{trust}} = 50.80$, $SE = 2.85$, $F(5, 596) = 0.83$, $p = .53$).

The second finding is that following pitfall information, learning about the benefits and performance of an autonomous driving algorithm restored ratings of trust (all $ps < .004$). However, this restoration was greater for performance information as compared to the benefits information, $M_{\Delta \text{ trust, performance} | \text{ pitfall}} = 12.11$ vs. $M_{\Delta \text{ trust, benefits} | \text{ pitfall}} = 7.36$, $F(1, 398) = 8.76$, $p = .003$. This trust restoration was evident in two places, as illustrated by Panel B of Figure 3.

First, trust was restored in information-ordering conditions where pitfall information was the first information parcel; that is, in two

⁴ Upon completing the main task, participants also completed the algorithm knowledge and awareness scale (Dogruel et al., 2022). Although we do not consider those data further here, we present the full data in the associated Open Science Framework archive listed after the main article.

Figure 3
Trust Ratings for Experiment 4



Note. Information-order condition is separated by colour in both panels. Panel A shows difference scores in trust ratings following each information parcel, where scores higher than 0 indicate a relative increase in trust, and scores lower than 0 indicate a decrease in trust. Individual dots represent participant-level difference scores, and boxplots show the median and central interquartile ranges. Panel B shows the mean trust scores, with standard errors for each information-order condition at each time point in the experiment. On the trust scale, values above 50 were indicative of a preference for the algorithm driver, values below 50 were indicative of a preference for the human driver, and values of precisely 50 indicated equal preference for both agents. Info. = information. See the online article for the color version of this figure.

conditions: where pitfall information was followed by benefits information (blue-color order, $M_{\text{increase}} = 7.18$, $SE = 1.53$) and where pitfall information was followed by performance comparison information (pink-color order, $M_{\text{increase}} = 14.50$, $SE = 1.84$). Second, trust was restored in information-ordering conditions where pitfall information was the second information parcel; that is, (i) where pitfall information was preceded by benefits, then followed by performance comparison information (brown-color order, $M_{\text{increase}} = 9.72$, $SE = 1.44$), and (ii) where pitfall information was preceded by performance comparison information, then followed by benefits information (turquoise-color order, $M_{\text{increase}} = 7.54$, $SE = 1.62$).

More broadly, our third finding was that despite these quantitative changes in trust, we did not find evidence that people changed their overall levels of trust in algorithms, when quantified by the difference between initial baseline trust ratings and final trust ratings (Figure 4, Panel A; H_1 : initial—last rating $\neq 0$ vs. H_0 : initial—last rating = 0, all $ps > .26$). Additionally, we failed to

detect a difference in overall trust between the information-ordering conditions, $F(1, 596) = 1.54$, $p = .17$. Taken together, our results suggest that despite local quantitative changes in response to individual parcels of information, people's overall preferences for an autonomous driving algorithm were difficult to shift.

To further visualize these sticky initial preferences, we classified participants ($N = 602$) depending on their preferred agent (algorithm, human, or equally preferred) at baseline (i.e., before any information was presented), as well as whether their preference shifted by the final rating relative to their baseline initial rating. We present a histogram of this classification in Figure 4, Panel B. Classification of the baseline and final trust ratings was conducted as follows: Rating responses greater than 50 were classified as *algorithm-preferred*, responses less than 50 were classified as *human-preferred*, and responses of exactly 50 were classified as *equal preference*. Overall, 79.3% of participants ($n = 477$) retained their initial preferences for either autonomous driving algorithm, human driver, or equal

preference for both (three left-most columns in Figure 4, Panel B). Only a small proportion of our participants (9.1%, $n = 55$) switched their preferences to the algorithmic agent or toward equal preference for both human and autonomous algorithmic drivers (2.8%, $n = 17$). Therefore, in summary, although we find evidence of local shifts in trust in response to algorithm information, for the vast majority of our participants, trusting an algorithm was a sticky—and for many, a tricky—thing.

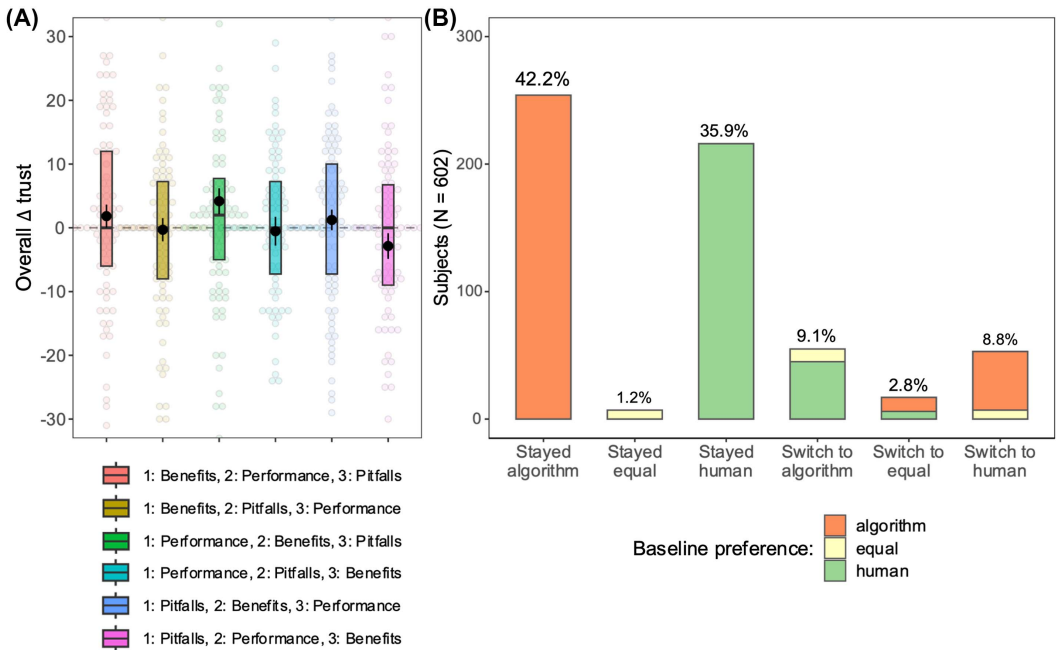
General Discussion

In four experiments, we sought to understand how algorithm information influenced people’s trust in algorithms. Across four real-world scenarios, we presented summaries of published academic studies in which an algorithm outperformed human agents in the same environment

and examined the role of information elements (e.g., degree of algorithm superiority, complexity of explanation about how an algorithm works) in shaping people’s trust in algorithms. We find that such descriptive summaries increased trust in algorithms, and, additionally, our within-participants extensions of Castelo et al. (2019) provided the granularity to identify for whom trust increased and to what degree. Promisingly, these results suggest that one avenue for improving attitudes toward algorithmic decision aids is to provide individuals with information about how algorithms have been validated (Dietvorst et al., 2015).

Our experiments provided three core advances to understanding trust in algorithms. First, alongside performance comparisons, our experiments support the idea that explanations of how an algorithm works increased trust in that algorithm. We find that the presentation of

Figure 4
Overall Changes in Trust Ratings and Preferences in Experiment 4



Note. Panel A shows differences between initial trust ratings and final trust ratings as a function of experimental condition. Faded individual dots represent individual participants, boxplot displays central interquartile range and median difference scores, and black dots show means and standard error. Horizontal dotted intercept line at $y = 0$ shows no difference. Information ordering condition is shown in color with the legend below both panels. Panel B shows the final preferences for each agent as a function of whether the individual stayed with their baseline preference or switched from their baseline preference. Baseline preference is indicated in color according to the legend within Panel B, and overall percentages are shown above each histogram bar. See the online article for the color version of this figure.

relatively complex—but importantly, also simple—explanations increased trust across our scenarios. For individuals who may be less familiar with software development, process explanations offer a different route to trust from that offered by performance comparisons. An explanation allows the user to conceptualize the machinations underlying their data and how the algorithm may be able to leverage external information to the user's advantage. This process of introducing new information is likely the reason why, across our experiments, trust increased the most in the medical scenario. Medical diagnosis requires such specialized knowledge that individuals might have been surprised to learn that algorithms can provide useful digital support systems. By comparison, the everyday familiarity with movie-recommender systems makes an explanation of the movie recommendation algorithm less informative and, thus, perhaps less persuasive.

Optimistically, our data suggest that the uptake of algorithmic recommendations in some domains may be only a matter of time. This is not to say, however, that algorithms are a panacea. For example, while there is a general sentiment within the medical field that decision aids are useful (e.g., [Graham et al., 2007](#); [Ridderikhoff & van Herk, 1999](#)), doctors who use decision aids are sometimes perceived by patients to be less skilled compared to their unaided or even human-aided counterparts ([Arkes et al., 2007](#); [Shaffer et al., 2013](#)). These trends have stood the test of time, albeit worth noting that the complexity of the decisions tools has considerably evolved in recent decades ([Sutton et al., 2020](#)). Our data shed light on the fact that providing information to decision-makers ameliorates a degree of distrust—at least locally within a specific context—which, in compound with time and experience, may encourage individuals to adopt algorithmic decision aids.

Our second contribution is demonstrating that not all algorithm-related information is weighted equally. Across experiments, we observed individual variation in how one's trust changes in response to each piece of algorithm information. For some individuals, learning about the performance validations from academic summaries was convincing, consistent with [Castelo et al. \(2019\)](#), while other individuals preferred explanations, consistent with [Dzindolet et al. \(2003\)](#). Indeed, one could find several examples

across the literature that report how each factor, in isolation, can offer a route to trusting algorithms. However, only by combining distinct factors together can we observe how decision-makers weigh and potentially trade off each piece of new information.

Our data make inroads into understanding how pitfalls, explanations, and performance evaluations about algorithms are evaluated against one another. Trade-offs are commonplace in reality, such as in medical contexts where people make difficult end-of-life decisions ([Waller et al., 2021](#)) or consumer decisions about different attributes of purchasing electronics ([Hawkins et al., 2014](#)). These situations require the decision-maker to weigh different attributes, akin to our performance summaries and explanations, and decide between a cumulative set of factors. In future work, one interesting avenue is to explore how these differences in preferences relate to algorithm behaviors. One potential route is to use discrete choice experiments ([Gronau et al., 2023](#)) that gauge the key attributes in people's decision making before offering actual options that differ across these attributes (e.g., use a bespoke movie recommender, a generic recommender, or select from a list). As a starting point, our work points toward movie-recommender algorithms as a familiar testbed by which behavioral outcomes and trust can be observed and elicited.

Our third key finding was that, despite local, quantitative effects on trust, overall algorithm-human preferences were difficult to shift. We found stability in agent preferences, particularly in Experiment 4, where multiple localized changes in response to different information resulted in minimal changes overall.⁵ In one sense, the stability of people's preferences is unsurprising—our subjects were presented with brief samples of information about an algorithm, which could have been overshadowed by their own experiences and prior beliefs. However, our contribution here is to place bounds on algorithm information effects (e.g., [Castelo et al., 2019](#); [Dzindolet et al., 2003](#)), where attempting to increase trust across different informational factors is evidently not additive. Instead, our experiments suggest individuals reach a ceiling at which they are

⁵ Examining the car scenario for Experiments 1, 2, and 3 reveals a similar stickiness of initial preferences. See [Supplemental Figure S3](#) for comparison.

not willing to shift their trust ratings in response to algorithm information.

An alternative explanation for sticky algorithm preferences is that participants may have been disengaged from the task; if this were the case, one might expect that a large proportion of participants never changed their ratings across the scenarios. However, one notable consistency across all four experiments is that the majority of participants—roughly 85% within each experiment—updated their trust ratings to at least one scenario/algorithm information summary (see [Supplemental Figure S4](#)). This majority leads us to believe that disengagement is unlikely the root of sticky algorithm preferences but rather that participants were not sufficiently swayed by the information.

Evidently, our information summaries were sparse. In our experiments, the description of the algorithm and the decision context were brief, and so individuals likely imported a larger set of prior expectations of each algorithm within the different scenarios. A mismatch between these expectations and information summaries would compromise trust in an algorithmic tool ([Himmelstein & Budescu, 2023](#)). For example, [Lehmann et al. \(2022\)](#) found that simple explanations of a supply–demand forecasting algorithm (e.g., operates by simple averaging) violated people’s expectations that such an algorithm should be more complex. Beyond this constraint, how to convince decision-makers when it is appropriate to trust algorithmic recommendations is an open question. Safety-critical domains, such as medical advice, would likely necessitate more extensive validations, aided by the credibility or endorsement of existing trustworthy institutions. In this sense, sticky preferences are likely beneficial and warranted until many endorsements of a new technology coalesce.

Algorithms may be here to stay for the long run, but in many domains, their application has just reached infancy. Trust in these tools depends on many factors, like the utility to the user, the development of more accurate predictions, and perhaps, undercurrent to all factors, the incentives and reputations of companies doing their development. Considered together, perhaps an apt comparison is that like when interacting with people, trust is tricky. However, our investigations make inroads to understanding the different kinds of algorithm information that shapes, breaks, and restores people’s trust in these tools.

References

- Angwin, J., & Larson, J. (2016). *Machine bias, risk assessments in criminal sentencing*. Propublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, 27(2), 189–202. <https://doi.org/10.1177/0272989X06297391>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Blanco, M., Atwood, J., Russell, S. M., Trimble, T., McClafferty, J. A., & Perez, M. A. (2016). *Automated vehicle crash rate comparison using naturalistic data*. Virginia Tech Transportation Institute.
- Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239. <https://doi.org/10.1002/bdm.2155>
- Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825. <https://doi.org/10.1177/0022243719851788>
- Dastin, J. (2018, October 11). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
- De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47, 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Dogruel, L., Masur, P., & Joeckel, S. (2022). Development and validation of an algorithm literacy scale for internet users. *Communication Methods and Measures*, 16(2), 115–133. <https://doi.org/10.1080/19312458.2021.1968361>
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies*, 58(6), 697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- Graham, I. D., Logan, J., Bennett, C. L., Presseau, J., O’Connor, A. M., Mitchell, S. L., Tetroe, J. M., Cranney, A., Hebert, P., & Aaron, S. D. (2007).

- Physicians' intentions and use of three patient decision aids. *BMC Medical Informatics and Decision Making*, 7(1), Article 20. <https://doi.org/10.1186/1472-6947-7-20>
- Gronau, Q. F., Bennett, M., Brown, S., Hawkins, G., & Eidels, A. (2023). Do choice tasks and rating scales elicit the same judgements? *Journal of Choice Modelling*, 49, Article 100437. <https://doi.org/10.1016/j.jocm.2023.100437>
- Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive Science*, 38(4), 701–735. <https://doi.org/10.1111/cogs.12094>
- Himmelstein, M., & Budescu, D. V. (2023). Preference for human or algorithmic forecasting advice does not predict if and how it is used. *Journal of Behavioral Decision Making*, 36(1), Article e2285. <https://doi.org/10.1002/bdm.2285>
- Krishnan, V., Narayanashetty, P. K., Nathan, M., Davies, R. T., & Konstan, J. A. (2008, October). *Who predicts better? - Results from an online study comparing humans and an online recommender system* [Conference session]. RecSys'08: Proceedings of the 2008 ACM conference on Recommender systems, New York, NY, United States.
- Lehmann, C. A., Haubitz, C. B., Fügner, A., & Thonemann, U. W. (2022). The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management*, 31(9), 3419–3434. <https://doi.org/10.1111/poms.13770>
- Liang, G., Sloane, J. F., Donkin, C., & Newell, B. R. (2022). Adapting to the algorithm: How accuracy comparisons promote the use of a decision aid. *Cognitive Research: Principles and Implications*, 7, Article 14. <https://doi.org/10.1186/s41235-022-00364-y>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Lohr, S. (2016, October 17). IBM is counting on its bet on Watson and paying big money for it. *The New York Times*. <https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html>
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. <https://doi.org/10.1037/11281-000>
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). *Manipulating and measuring model interpretability* [Conference session]. CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan. ACM. <https://doi.org/10.1145/3411764.3445315>
- Rago, A., Cocarascu, O., Bechliyanidis, C., Lagnado, D., & Toni, F. (2021). Argumentative explanations for interactive recommendations. *Artificial Intelligence*, 296, Article 103506. <https://doi.org/10.1016/j.artint.2021.103506>
- R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ridderikhoff, J., & van Herk, B. (1999). Who is afraid of the system? Doctors' attitude towards diagnostic systems. *International Journal of Medical Informatics*, 53(1), 91–100. [https://doi.org/10.1016/S1386-5056\(98\)00145-2](https://doi.org/10.1016/S1386-5056(98)00145-2)
- Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4), 260–278. <https://doi.org/10.1080/12460125.2020.1819094>
- Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making*, 33(1), 108–118. <https://doi.org/10.1177/0272989X12453501>
- Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146, Article 102551. <https://doi.org/10.1016/j.ijhcs.2020.102551>
- Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3, Article 17. <https://doi.org/10.1038/s41746-020-0221-y>
- Waller, A., Wall, L., Mackenzie, L., Brown, S. D., Tattersall, M. H., & Sanson-Fisher, R. (2021). Preferences for life expectancy discussions following diagnosis with a life-threatening illness: A discrete choice experiment. *Supportive Care in Cancer*, 29(1), 417–425. <https://doi.org/10.1007/s00520-020-05498-7>
- Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4), 403–414. <https://doi.org/10.1002/bdm.2118>

(Appendix follows)

Appendix

Algorithm Information Statements

We present the information statements for each scenario below. Note that the autonomous driving car scenario used across all experiments has two additional statements included in Experiment 4. For Experiment 1, the performance evaluation statements included a numeric percentage in bold that was doubled in the “double %” condition. For Experiment 2, the simple explanations were contrasted against the complex explanations. For Experiment 3, the standard undoubled performance % summary in the performance evaluation statement was presented alongside the complex explanation of the algorithm. The initial statement below each scenario was presented in all experiments before any information parcel.

Car Driving Scenario

A recent study conducted by professional academic researchers showed that cars driven by algorithms experience fewer accidents compared to human drivers.

Prompt: Who would you trust more to drive a car safely?

Performance evaluation: In the study, cars driven by algorithms were 28% safer than cars driven by humans.

Simple: When the algorithm detects danger, it can brake faster than a human driver can.

Complex: The algorithm used a system of cameras to predict if an object in the environment would collide with the car. If there was the possibility of danger, the algorithm calculated the object's distance and took precautionary action (e.g., braking, or changing lanes) faster than a human driver could.

Benefits: Why it is better—The systems built into self-driving vehicles have eliminated the most common types of human driving errors that have led to tragic outcomes such as driveway reversing accidents and mistaken brake/accelerator pedals while parking.

Pitfalls: Potential pitfalls—Although comprehensively trained, the algorithms for recognizing braking situations cannot cover all contingencies on the road. For example, odd scenarios such as semitrailers blocking multiple lanes of a highway while turning, pedestrians walking their bikes rather than riding them, and kangaroo movement

patterns have previously led to fatal outcomes because algorithms failed to predict the object's movement and distance from the vehicle.

Movie Recommendation Scenario

A recent study conducted by professional academic researchers showed that an algorithm can predict what movies people like with higher accuracy than other people.

Prompt: Who would you trust more to recommend you a good movie?

Performance evaluation: In the study, the algorithm predicted what movies people liked with 20% more accuracy than other movie-watchers.

Simple: The algorithm recommended the most popular movie in cinemas at the time.

Complex: The algorithm asked the person to enter their top three movies. The algorithm gathered information about the person's preferred genre as well as actors/actresses. It then matched the person's preferences against thousands of people with similar preferences to recommend the highest-rated movie.

Joke-Funniness Prediction Scenario

A recent study conducted by professional academic researchers showed that an algorithm can predict how funny someone will find several jokes with higher accuracy than a person's close friend.

Prompt: Who would you trust more to predict how funny you would find a joke?

Performance evaluation: In the study, the algorithm predicted how funny someone found a joke with 61% accuracy, whereas the person's close friend only predicted with 54% accuracy. The algorithm was therefore about 7% more accurate than the person's own friend.

Simple: The algorithm uses an average funniness rating from the last 10 people who heard that joke.

Complex: The person first rated how funny they found a dozen jokes. The algorithm then matched the person to other users in a database with similar ratings and created a “humor” profile. Joke ratings are generated by matching not only the highest ratings from a database of

(Appendix continues)

thousands of jokes but also the person's specific humor profile.

Medical Scenario

A recent study conducted by professional academic researchers showed that an algorithm can recommend a treatment plan for cancer better than a human doctor.

Prompt: Who would you trust more to recommend you a cancer treatment plan?

Performance evaluation: In the study, the algorithm was able to recommend the same treatment plan as human doctors in 990 out of 1,000 cases and, in 30% of cases, identified treatment options that the human doctors had missed.

Simple: The algorithm recommends a newly developed treatment plan that most doctors were unaware of.

Complex: The algorithm was fed the patient's medical history and current symptoms. It then identified similar patients treated at the hospital and the most successful treatment plan for those symptoms. This treatment plan is then cross-referenced against patients from other hospitals around the country. If they were given a different treatment plan, the algorithm presented both options to the doctor.

Received May 30, 2023

Revision received December 10, 2023

Accepted December 20, 2023 ■

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <https://my.apa.org/portal/alerts/> and you will be notified by e-mail when issues of interest to you become available!