Trusting algorithms: performance, accuracy, and sticky preferences

Garston Liang & **Ben R. Newell**
School of Psychology, UNSW, Sydney, Kensington
Sydney, NSW, Australia, 2052

Abstract

What information guides individuals to trust an algorithm? We examine this question across three experiments that consistently found *explanations* and *relative performance* information increased trust in an algorithm relative to a human expert. Strikingly however, in only 23% of responses (414/1800) did an individual's preferred agent for a task (e.g., driving a car) change from human to algorithm. Thus, initial preferences were 'sticky' and largely resistant to large shifts in trust. We discuss theoretical and practical implications of this work and identify important contributions to our understanding of how summaries of information can improve people's willingness to trust decision aid algorithms.

Keywords: algorithm; accuracy; expert

Introduction

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. Algorithms also extend beyond the mundane to assist doctors in medical imaging, assess risks of recidivism for bail sentencing, and pilot autonomous vehicles. Essentially, if the data exists, machine learning algorithms can learn to make predictions. However, whether these predictions are recruited or ignored depends on convincing decision-makers of their merits.

In this article, we investigate what information convinces individuals to trust an algorithm. Across three experiments, we presented real-life vignettes in which participants indicated their preferences for an algorithm or human agent to perform a task (e.g., drive a vehicle). We paired these vignettes with information from academic studies where machine learning algorithms outperformed a human counterpart. To preface the results, we show that by including information about a) the degree of algorithm improvement, and b) how the algorithm operates, increases people's trust in the algorithm. These results point to the kinds of information interventions that may encourage individuals to consider recruiting 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 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, Simmons, & Massey, 2015). Recent attention has focused on how different information about the algorithm affects this degree of aversion. For example, when individuals were given positive descriptions of the algorithm's capabilities prior to any direct experience, they held favourable attitudes towards using its recommendations (Logg, Minson, & Moore, 2019).

Performance information

To help a person decide to use an algorithm, a key piece of information might be how the algorithm's performance compares to a human operator. Along this vein, Castelo, Bos, and Lehmann (2019) created summaries of nine previous experiments that directly compared algorithms to human agents. Each summary provided empirical evidence of different scenarios in which algorithms outperformed human agents ranging from movie recommendation 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 (Shah, Mullainathan, & Kleinberg, 2019). As a separate experiment, Castelo et al. then presented these nine summaries to participants who then indicate 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 affected people's trust. Presumably, studies that showed larger improvements with an algorithm provide the grounds for greater trust in its capabilities. However, Castelo et al. could only indirectly assess this prediction across qualitatively different scenarios, e.g., comparing the 7% algorithm improvement for joke-telling to 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

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

(provided you asked nicely). The same cannot always be said of machine learning systems. Between the large training datasets, hidden layers in the model's architecture, and the complex underlying statistical methods, it is entirely possible that some of the reluctance people experience with algorithm aversion 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, ethics of a decision process, or the alignment between the user's actual needs to the algorithm's optimisation function (Doshi-Velez, Kim, 2017). These broad concerns have sprouted new subfields of research in Explainable Artificial Intelligence (XAI) and prompted wide-reaching protection laws such as the EU's General Data Protection Regulations (Shin, 2021). Our experiments explore this core idea that trust is founded upon transparency.

We directly compare simple explanations to more complex versions to understand how explanatory detail may play a role. Previous work by Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) found that even simple explanations about why an algorithmic aid could occasionally 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. Indeed, much resistance within clinical circles to the statistical decision rules advocated by Meehl (1954) was centred 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 given recommendation (Rago et al., 2021). We tested this possibility by presenting both simple and more complex explanations of the algorithm's process.

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 that 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 to differing degrees of algorithm superiority over the human expert (e.g., 20% better vs. 40% better) and to explanations of how the algorithm operates.

In Experiment 1, we investigated whether doubling the performance improvement information further increased trust in an algorithm. In Experiment 2, we contrasted simple and complex explanations of how the algorithm operated. In Experiment 3, we presented both pieces of information in a counterbalanced order. We hypothesised that a) larger degrees of superiority in favour of the algorithm would lead to larger increases in trust ratings, b) complex explanations would lead to greater increases in trusts than simple explanations, and c) that compounding both parcels of information would lead to a sub-additive ceiling on 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 Methods

We describe the methods for all three experiments as each follow the same general procedure.

Participants

In each experiment, we recruited 150 participants (N = 450) 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 (N_{male} = 136, N_{female} = 299, $N_{\text{other gender}}$ = 12, $N_{\text{prefer not to say}}$ = 1; M_{age} = 27.9). Relative to Castelo et al, the within-subject design coupled with larger sample sizes provided additional power to detect algorithm-information effects. Participants were renumerated £1.66 for completing the 10-minute experiment.

Materials

Trust ratings The experiment was programmed in Javascript using JSPsych (de Leeuw, 2015) and completed on a webpage. The task required participants to input two ratings on a slider scale to the question "Who would you trust more to do [X]?" in the respective scenario. The scale spanned from 0 to 100 with labels for algorithm (0), equal trust in algorithm/human (50) and human expert (100). For the *initial trust rating*, 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 the *post-information rating*, the start point of the slider was set to the individual's initial rating and did not require any interaction to proceed if they wished to retain that rating.

Algorithm Scenarios & Information

In each experiment, participants were presented four scenarios that were selected from Castelo et al. (2019) in a randomized order. These four scenarios were 1) recommending a treatment plan for cancer, 2) safety when driving a car, 3) rating the funniness of a joke, and 4) recommending a movie. Although tangential to our pursuits, we used Castelo et al.'s objectivity ratings[1] to guide our selection of the two highest (cancer, car) and two lowest

[1] In a separate study, Castelo et al., (2019) asked participants to rate a given task, e.g., providing movie recommendations, as objective/subjective on a scale ranging from 1-100, 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.

objectivity-rated scenarios (joke-telling, movie). Our intent was to identify a range of situations where initial preferences for algorithmic decision making may be more favourable (self-driving vehicles & movie recommendation) compared to areas traditionally reserved for human judgement (joke-telling & medical diagnosis).

Within each scenario, a text summary was presented after the initial trust rating that summarized the results of a published study comparing an algorithm to a human expert. We present all the text summaries in the appendix. Here we illustrate with an example of the performance information for the movie scenario presented in Exp. 1 & Exp. 3:

"A recent study conducted by professional academic researchers showed that an algorithm can predict what movies people will like with 20% more accuracy than other movie-watchers.

Who would you trust more to recommend a good movie?"

The prompt then asked participants to input a *post-information* rating along the same slider scale if they wished to update their *initial rating*. The above performance text summary and prompt were identical to those in Castelo et al. (2019) that demonstrated an algorithm information effect in a between subject's design. Our within-subjects 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.

Design

All three experiments used a two between-by-within design. Scenario and ratings were within-subject factors. The between subject factor differed by experiment. Exp. 1 manipulated the stated degree of improvement of the algorithm (e.g., the *20% more accurate* in the movie scenario), presenting either a veridical percentage (actual %) or doubled percentage. Exp. 2 manipulated the explanation complexity related to how the algorithm generated its recommendation (simple vs. complex[2]). Exp. 3 combined the veridical improvement information from Exp. 1, and the complex explanations from Exp 2. and manipulated the order of information presentation between-subjects (accuracy-first, explanation-first). See Figure 1 for a flow diagram comparing the designs.

Procedure

Participants were instructed that their task was to indicate their trust in either an algorithm or human judge across four different scenarios where the order of the four scenarios was randomised. All participants entered an initial trust rating to the prompt "who would you trust more [in scenario X]?". In Exp. 1 and Exp 2., the *initial rating* was made in the absence of the study summary. In Exp. 3, the *initial rating* was made with study summaries present.

The conditions diverged based on the content of the study summaries. Once inputted, study summaries appeared on the next page above the same slider scale, where the start point was set to their initial rating. Individuals read through the summary and were asked "Who would you trust more to [in scenario X]? Move the slider now if you'd like to update your previous response." Each subject provided two ratings for all four scenarios which upon completion, a free-response question asked individuals how they approached the task and whether they felt their responses were influenced by algorithm information.

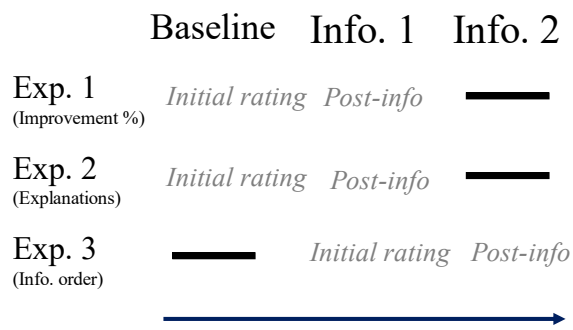

Figure 1: Diagram illustrating the designs. Long arrow indicates the experiment progression with large dashes indicating absent ratings for a given experiment.

Results

Analysis was conducted in R (RCore 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 for the degree of trust change and report the means and standard errors with each analysis. Follow up tests use Bonferroni corrected p-values.

We organise the results as follows: first, we analyse the initial trust ratings in Exp. 1 and Exp. 2 followed by the degree of trust change after the presentation of performance information (Exp. 1) and explanations (Exp. 2). To preface the results, we demonstrate that mean trust in the algorithm increased in every scenario following the presentation of algorithm information. Next, we analyse the data for Exp. 3 to show the order of information does not affect mean trust ratings.

Although we consistently find a shift towards the algorithm in the aggregate, cross-experiment analyses show that many individuals did not change their trust ratings. We present data for the proportion of responses where individuals did not change their trust ratings and conclude by considering our data in the broader context of convincing individuals to accept and trust an algorithmic agent.

[2] Piloting (N = 100) confirmed that subjects overwhelmingly perceived the difference in explanation complexity.

Initial trust (Exp. 1 & Exp.2)

Trust in algorithms was highest in the "movie-recommendation" scenario (M = 48.6, se = 1.9; first-red column in Figure 2) nearing equal trust for both human experts and algorithms. This was significantly higher than the trust in algorithms for all other scenarios where we find a general preference for human experts, χ^2 (3) = 40.27, p < 0.001. Next, we calculate the mean trust change scores by subtracting the *initial ratings* from *post information* ratings and present these in Figure 3.

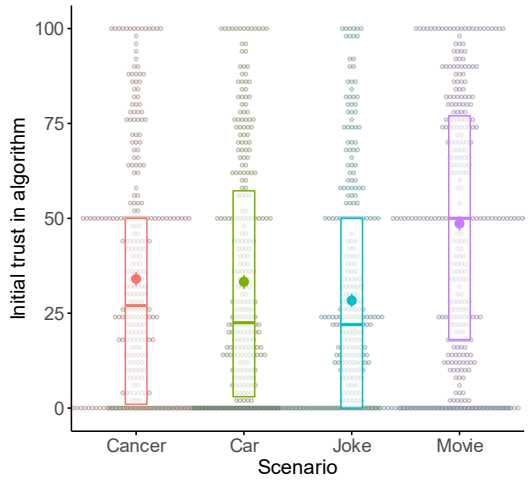

Figure 2. Initial trust ratings in the algorithm as a function of scenario. Means and standard errors plotted alongside boxplot of median and central interquartile range. Small dots represent raw data. Scale ranged from 0 (trust a human expert) to 100 (trust an algorithm) with the 50-midpoint representing equal trust in human/algorithm. Note Exp. 1 & Exp. 2 combined in Figure 1 as ratings were made prior to conditions diverging.

Information effect (Exp. 1 & 2)

In Exp. 1, trust in the algorithm increased in all scenarios and conditions (two left-most panels in Figure 3; all scenario p's > 0.004 vs. M $_{diff}$ = 0). Notably the stated accuracy of the algorithm did not affect the increase in trust ratings ($M_{doubled\ \%}$ = 12.8 vs. $M_{actual\ \%}$ = 15.4%; χ^2 (1) = 1.90, p = 0.17). Trust in the algorithm increased the most in the cancer diagnosis scenario (M = 20.5, se = 2.16) compared to all other scenarios (χ^2 (3) = 34.36, p < 0.001). The scenario by stated accuracy interaction was likely driven by trust ratings in the movie scenario that were significantly lower than all scenarios in the actual-% condition (all p's < 0.005) but did not differ from other scenarios in the doubled-% condition (all p's > 0.14; χ^2 (3) = 9.31, p = 0.03).

In Exp. 2, trust in the algorithm increased in all scenarios and conditions (two right-most panels in Figure 3; all scenario p's > 0.003 vs. M $_{diff}$ = 0). Contrary to our expectations, the complex explanations did not inspire greater increases in trust compared to the simple – and more transparently flawed – explanations ($M_{complex}$ = 11.0, se, 1.07 vs. M_{simple} = 12.7, se = 1.23; χ^2 (1) = 0.85; p = 0.36). Again, we observed the greatest increase in trust for the algorithm in the cancer scenario (M = 16.3, se = 1.75) with the lowest increases in trust in the movie-recommendation scenario (M = 7.75, se = 1.35; t (453) = 3.94, p < 0.001). A tentative explanation for this pattern is that trust was initially higher for the movie-recommendation scenario and so participants may have already been convinced prior to additional information reinforcing the superiority of an algorithm. All other effects were non-significant.

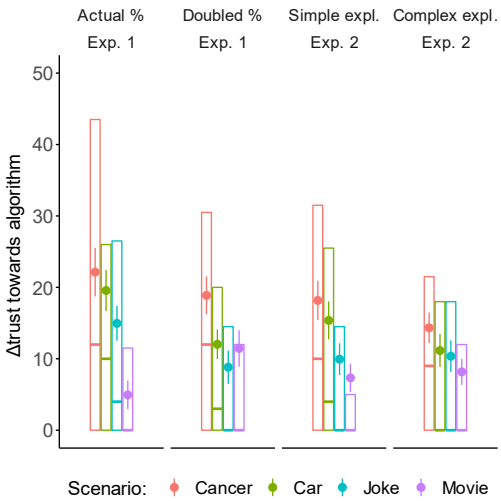

Figure 3. Mean change in trust ratings as a function of scenario and experimental condition in Exp. 1 & 2. Standard errors shown alongside boxplots of the central interquartile range. Colour indicates scenario, panels indicate condition.

Order of algorithm information (Exp. 3)

The promising results of Exp. 1 & Exp. 2 suggests that both performance information and explanation statements can increase trust in algorithms in a range of situations. In Exp. 3, we presented the complex explanations and the actual accuracy improvement information and manipulated the order of presentation. We expected that the improvement-first condition to have smaller increases in trust because after one knows the algorithm is superior, subsequently learning the explanation may only be a curiosity. We present the initial ratings by information condition in Figure 4.

Initial trust ratings were highest for the movie-recommendation scenario (M = 63.2, se = 2.27) and lowest for the joke-rating scenario (M = 39.1, se = 2.42). The type of information also interacted with the scenario where trust ratings were lower in the movie-scenario with improvement information ($M_{explain}$ = 65.9 vs. $M_{improve}$ = 60.5, t (147) = 1.18, p = 0.24) but higher with improvement information in all other scenarios (cancer scenario $M_{explain}$ = 51.6 vs. $M_{improve}$ = 58.7, t (147) = 1.57, p = 0.12; car scenario $M_{explain}$ = 39.7 vs. $M_{improve}$ = 52.0, t (145) = 2.51, p = 0.12; joke scenario $M_{explain}$ = 37.6 vs. $M_{improve}$ = 40.6 t (147) = 0.62, p = 0.54). These initial results suggest individuals may find improvement information more convincing for trusting algorithms rather than explanations of how they work. Next, we compare how

the additional algorithm information changed people's trust in an algorithm.

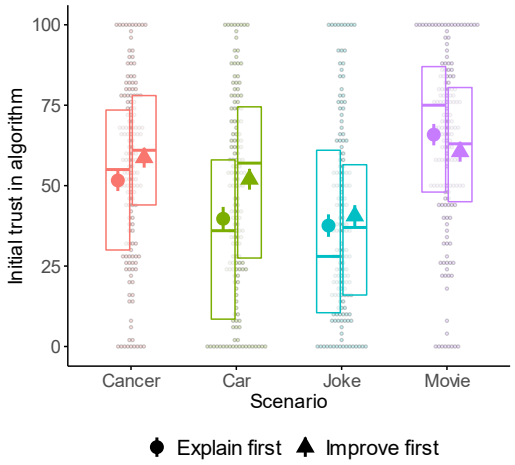

Figure 4. Initial trust ratings in the algorithm as a function of scenario and information condition indicated by shape. Means and standard errors plotted alongside boxplot of median. Small dots represent raw data. Scale ranged from 0 (trust a human expert) to 100 (trust an algorithm) with the 50-midpoint representing equal trust in human/algorithm.

Information effect (Exp. 3)

In Exp. 3, additional algorithm information raised the trust ratings in all scenarios (all p's < 0.04) except for the movie-recommendation scenario with improvement information first (furthest right column, Figure 5, M = 4.07, se = 2.09; p = 0.06). At least numerically, we observed that the greatest increase in trust again in the cancer diagnosis scenario (M = 8.97, se = 1.30) and lowest in the movie recommendation scenario (M = 4.69, se = 1.31). However, we failed to observe significant differences between the scenarios χ^2 (3) = 7.02, p = 0.07), information order conditions (χ^2 (1) = 2.47, p = 0.12), or an interaction (χ^2 (3) = 1.14, p = 0.77). The notably smaller increases in algorithm trust are likely related to the fact the additional information in the initial rating raised the 'baseline' level of trust. We explore this idea further in the General Discussion.

General Discussion

In three experiments, we sought to understand what algorithm information influenced people's trust in algorithms. Across a range of scenarios, we presented different summaries of published academic studies in which an algorithm outperformed human agents in the same environment. We consistently find that such study summaries increased trust in algorithms. Promisingly, these results suggest that one avenue for improving attitudes towards algorithmic decision aids is to provide individuals with information about how algorithms have been validated or explain how they might work (Dietvorst, Simmons, & Massey, 2014).

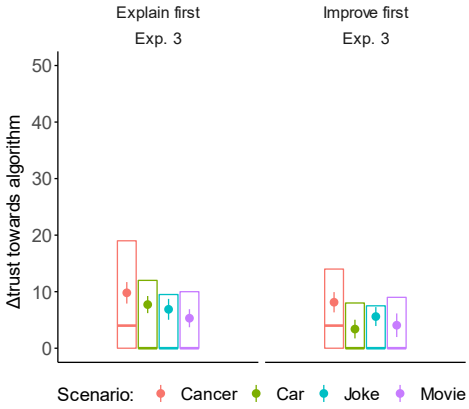

Figure 5. Mean change in trust ratings as a function of scenario and experimental condition in Exp. 3. Standard errors shown alongside boxplots of the central interquartile range. Colour indicates scenario, panels indicate condition.

Our within-subjects extension of the original information effect reported in Castelo et al. (2019) provides a number of theoretically and practically important distinctions about how algorithm information increases trust. Firstly, our experiments provide the granularity to identify that not all individuals find the algorithm information useful. In other words, there were a number of individuals who did not change their *initial ratings* after the presentation of the algorithm information; an observation that is absent in Castelo et al's original between-subjects experiment. These proportions of non-movers varied across the scenarios though notably were consistently the lowest in the cancer scenario across the experiments (see Table 1). One tentative interpretation of these data is that, compared against the other scenarios, participants were least familiar with the procedures involved in cancer diagnosis. Presumably, this degree of relative ignorance meant that any algorithm-information provided sufficient reason to increase their trust ratings (cf. Bonezzi, Ostinelli, & Melzer, in print). Speaking to this interpretation is that the proportion of non-movers was typically highest in the movie-recommendation scenario, likely reflecting the familiarity with these recommender systems in daily life and so, corresponding high level of initial trust. In the absence of explicit tests of this knowledge, our interpretation remains speculative. However, our findings provide the foundations for distinguishing distinct drivers of the algorithm information effect in each scenario (Castelo et al., 2019). Taken together, our within-subjects design allows us to delineate a general increase in trust at the aggregate-level as comprised of a mixture of individuals thinking about the details of each scenario (Chen, Regenwetter, Davis-Stober, 2021)[3].

Optimistically, our data suggests that in certain domains, the uptake of algorithms face fewer barriers than others. This

[3] The proportion of non-movers in each experiment does not appear to be driven by a consistent proportion of obstinate responders. Subject-level analyses indicate that 84% of participants changed their trust ratings at least once across the four scenarios.

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 that use decision aids are sometimes perceived by patients to be less skilled compared to their unaided or even human-aided counterparts (Arkes, Shaffer, & Medow, 2007; Shaffer et al., 2013). Our data shed light on the fact that providing information to decision makers appears to ameliorate a degree of the distrust and encourage individuals to use helpful decision aids.

Table 1: Proportion of participants who did not change trust ratings following receipt of information (non-movers) in each scenario and experiment. Note that lowest proportion of non-movers (i.e., most convinced by information) in bold.

			Scenarios	
Experiment	Cancer	Car	Joke	Movie
1	**0.37**	**0.37**	0.48	0.55
2	**0.34**	0.49	0.48	0.62
3	**0.41**	0.48	0.57	0.47

The second extension of our experiments is that alongside performance information, we established explanations providing compelling reasons to trust an algorithm. We find that both simple and complex explanations increased trust in different algorithms. Furthermore, the initial ratings in Exp. 3 allow for a comparison of the relative persuasiveness of explanations against performance information. It appeared that performance information summaries were generally more persuasive, albeit with a nuance; explanations may be preferred if decision makers already trust the algorithm. This may be because an explanation provides new information; one who already trusts an algorithm may already know that it can surpass the performance of a human. Only an explanation provides novel information about how the algorithm functions and what potential pitfalls and benefits it can bring.

Lastly, our data place bounds on the algorithm information effect. Presenting both parcels of information did not lead to an additive increase in trust but rather seems to indicate individuals have a ceiling to which are willing to shift their ratings. With regards to broader implications, it may be prudent to consider a simple statistic. In only 23% of responses did an individual's preferred agent change (414/1800, see first three columns of Figure 6). If we imagine participants treated our continuous scale as a binary choice between a human or algorithmic agent, then evidently, there are limits to how convincing short study summaries can be. In this sense, participants exhibit a 'sticky' initial preference that may only shift within the given range of the preferred agent. This bounded stickiness acknowledges that the uptake of algorithms is necessarily guided by more than brief bursts of information. The information in our experiments were admittedly brief but provide a comprehensive demonstration of the capacity of decision makers to trust algorithms when convinced of their merits.

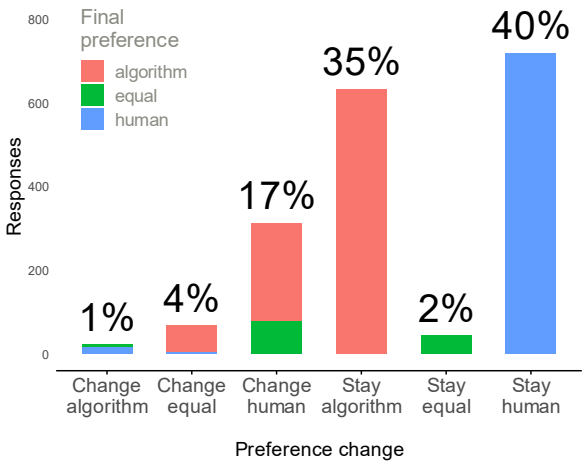

Figure 6. Proportion of responses where the initial preference changed (i.e., moved past the 50-mid point). First three columns indicate change; last three columns indicate retention. Colour indicates the final preference.

Conclusions

In three experiments, we explored the bounds of an algorithm information effect. We find that trust in algorithms can be increased by presentations of performance information *and* process explanations of the algorithm. Parceled together, we find that participants exhibit a stickiness in their initial preferences. Undoubtably, machine learning is a powerful tool. However, the more ambitious challenge is to decide which data, and consequently cultural norms and values encapsulated therein, we wish to guide the training of machine learning and indeed society at large. That challenge remains a uniquely human pursuit.

Appendix

For brevity, we have condensed the text summaries to the core accuracy information and note the references were not shown to participants. All algorithm information summaries began with the phrase: *A recent study conducted by professional academic researchers showed [X …]*

Cancer: … in **30% of cases** the algorithm identified treatment options missed by the human doctor (*Lohr, NYTimes 2016)* -- **Car**: … cars driven by algorithms were **28% safer** … (*Blanco et al., 2016)* -- **Joke**: … can predict how funny a person will find a joke with **7% more accuracy** than the person's own friend (*Yeomans et al., 2019*) -- **Movie**: … can predict what movies people will like with **20% more accuracy** than other movie-watchers. *(Krishnan et al.,2008)*

Again, for brevity, we present a condensed example of the explanations for the movie scenario.

Simple: The algorithm recommended the most popular movie at the cinemas. -- **Complex**: The algorithm asks the person to enter their top three movies … favourite genre, actors/actresses, and then matches those preferences against thousands of people …

Acknowledgements

We thank A. Prof. Castelo for kindly sharing the original data.

References

Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, ***27***, 189-202.

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**, 1–48.

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.

Bonezzi, A., Ostinelli, M., & Melzner, J. (in print). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of Experimental Psychology: General*.

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***, 220-239.

Chen, M., Regenwetter, M., & Davis-Stober, C. P. (2021). Collective Choice May Tell Nothing About Anyone's Individual Preferences. *Decision Analysis*, ***18***, 1-24.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, ***56***, 809-825.

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***, 114-126.

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***, 697-718.

Graham, I. D., Logan, J., Bennett, C. L., Presseau, J., O'Connor, A. M., Mitchell, S. L., ... & Aaron, S. D. (2007). Physicians' intentions and use of three patient decision aids. *BMC Medical Informatics and Decision Making*, ***7***, 1-10.

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. In *Proceedings of the 2008 ACM conference on Recommender systems* (pp. 211-218).

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

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.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. University of Minnesota Press.

Rago, A., Cocarascu, O., Bechlivanidis, C., Lagnado, D., & Toni, F. (2021). Argumentative explanations for interactive recommendations. *Artificial Intelligence*, ***296***, 103506.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL 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***, 91-100.

Shaffer, V. A., & Zikmund-Fisher, B. J. (2013). All stories are not alike: a purpose-, content-, and valence-based taxonomy of patient narratives in decision aids. *Medical Decision Making*, ***33***, 4-13.

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***, 102551.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, ***32***, 403-414.