Increasing Algorithm Appreciation in AI-based Decision Support Systems through Encouraging Theory of Machine

Katharina Olejnikov (4402014)

Social, Health, and Organisational Psychology

Utrecht University

Master Thesis

Supervisor: Baptist Liefooghe

Second Reviewer: Ruud Custers

UU-SER approval number: 22-0439

The manuscript should be made publicly accessible

24 June 2022

Abstract

While the use of artificial intelligence (AI) for decision-making is widespread, the technology cannot be fully realized when the end-user mistrusts it. To increase algorithm appreciation, literature supports the idea of clarifying how AI works. Rather than presenting AI as a black box, the framework Theory of Machine proposes an approach to explain artificial intelligence to end-users by contrasting it to human thinking. The present research examines the effects of Theory of Machine priming on algorithm appreciation. One hundred twenty-eight participants were randomly assigned to a priming condition where artificial intelligence was introduced as in Theory of Machine or as a black box. Namely, comparing human to algorithmic judgment or by only giving technical descriptions of AI's reasoning, respectively. Afterwards, participants performed an age guessing game where an algorithm aided as a decision support system. The extent to which participants aligned their answers to the algorithm's advice was used as a measurement for algorithm appreciation (weight of advice). Additionally, task difficulty was manipulated to explore possible moderation effects. Based on previous literature it was hypothesized that a Theory of Machine framing will increase algorithm appreciation compared to the Black Box framing. The hypothesis was not confirmed. The results showed no significant difference between the means of the weight on advice score on the framing conditions or task difficulty. We conclude that a Theory of Machine framing does not influence algorithm appreciation. Several explanations for this effect, limitations of the study, and suggestions for further research on how to increase trust in AI are considered.

Keywords: artificial intelligence, algorithm appreciation, algorithm aversion, trust, theory of machine, decision making

Increasing Algorithm Appreciation in AI-based Decision Support Systems through Encouraging Theory of Machine

Today Artificial Intelligence (AI) assists in many decision-making tasks, ranging from trivial choices, such as which movie to watch on Netflix, to more crucial judgments, such as in medical diagnoses (Logg, 2021). AI can be understood as an algorithm performing tasks which supposedly require human cognition (Fügener et al., 2021). Even though this process is common in many domains of everyday life, AI's reasoning lacks transparency for most individuals and is often perceived as a black box (Mahmud et al., 2022; Cadario et al., 2021; Logg, 2021; de Fine Licht & de Fine Licht, 2020; Watson et al., 2019). Therefore, people react in different ways to AI-based decision-support systems (Burton et al., 2019; Logg, 2019). On the one hand, people show algorithm appreciation, when endorsing suggestions from AI. On the other hand, people can also react with algorithm aversion, thus reject AI.

Algorithm aversion poses a problem in the application of AI in daily life. Even though AI often better predicts results than humans, in many cases people do not comply with AI's results (Kuncel et al., 2013; Fildes et al., 2007). When people do not collaborate with algorithm's advice, the technology cannot be fully realized (Logg, 2021). Logg (2021) describes this as the 'Last Mile Problem': AI has to be understood by the humans working with the algorithm to close the communication gap between AI's analysis and humans' application of those insights. In contrast, usually, the focus lies on engineers improving algorithms to enhance accurate results. Adapting human factors, as in how to overcome the last mile problem, opens a new stream of research outside of emerging AI's technology. Yeomans and colleagues (2019) propose to investigate beyond the 'what', AI's output, which is considering how the output is presented to ensure the end-user complies with AI's results. If we better understand what leads people to distrust AI, we can challenge AI's implementation and overcome algorithm aversion.

Previous studies tended to focus on how people react with algorithm aversion, especially once AI errs or is just perceived to err, despite the fact that AI outperforms human predictions in general (Burton et al., 2019; Dietvorst et al., 2015; Manzey et al., 2012; Dzindolet et al., 2002). Humans expect algorithms to be perfect and are disappointed when their expectations are not met, thus react with distrust. However, the goal of an AI-based decision aid is not to be perfect, but to make merely better decisions than humans. To overcome algorithm aversion in this scenario, potential solutions include end-users modifying the algorithm's output. Specifically, giving the end-user control over the outcome by letting them adjust AI's judgment simply by 0.1% increases algorithm appreciation (Dzindolet et al.,

2002). A major drawback is that this intervention hurts AI's more precise output, as the end-user adjusts it, to increase acceptance of algorithmic judgment in the end-user. Other approaches to overcome the last-mile problem are for instance matching the AI's and end-users' decision-making style (Westin et al., 2015). However, this implies either changing the algorithm, so that AIs are constructed according to the end-user's decision marking style. Or, that individuals can work with only certain AIs matching their decision-making style. Consequently, the technologies full potential cannot be realized.

Humans also how aversion when AI's judgment is accurate. Two recent systematic literature reviews on this subject found that even when humans and AI give identical advice, the one from humans is preferred over the same one from AI (Mahmud et al., 2022; Burton et al., 2019). Recent research proposes that AI does not only need to be accurate to be accepted by humans, but the procedure behind algorithmic judgment also needs to be understood by the end-user. Yeomans and colleagues (2019) argue this might be due to humans being unable to compare algorithmic processing to human judgment and their need to understand the AI system. To overcome aversion towards algorithms the end-users need to gain a better understanding of algorithms. Aversion could be caused by prior beliefs and pre-existing false expectations of algorithms functioning the interaction with those (Burton et al., 2019; Goodyear et al., 2016). In their review, Burton et al. (2019) found that training in algorithmic literacy namely, 'how to interact with algorithmic tools, how to interpret statistical outputs, and how to appreciate the utility of decision aids' (Burton et al., 2019, p.3) seems not sufficient to overcome algorithm aversion.

Interventions aiming to overcome algorithmic aversion in AI-based decision support focus on clarifying how AI makes decisions compared to human reasoning, might be the solution. Goodwin and colleagues (2013) show that when individuals are offered an explanation of how an algorithm works, it increases their stated trust in forecasting support system. In their research, users initially had a poor understanding of how the system draws a conclusion and providing them with an explanation about the forecasting support system increased trust in AI and collaboration. A similar study was conducted by Cadario and colleagues (2021) in the context of medical AI. Initially, participants exhibited a greater illusory subjective understanding of human decision-making compared to AI decision-making. As a consequence, participants preferred advice from humans. When providing explanations for AI decision-making, just the belief of laymen to understand algorithmic decision processes led to an increased algorithm appreciation and lowered resistance to cooperation (Cadario et al., 2021). All in all, a growing body of literature demonstrates how

clarifying differences between algorithmic and human decision making seems like a plausible solution to overcome algorithm aversion.

The aforementioned studies thus indicate that clarifying the algorithm's approach in decision making to the end-user may increase trust and acceptance in AI decision making. Users need to collaborate with AI's advice and yet have often little insight into how AI processes data ('thinks') and gives outputs ('judges') (Cadario et al., 2021; Logg, 2021; Burton et al., 2019; Yeomans et al., 2019). Research by Goodwin and colleagues (2013) and Cadario and colleagues (2021) mentioned earlier indicates establishing better technical know-how in end-users about algorithmic data processes improves compliance with advice in that user. Logg (2021) goes a step further with her Theory of Machine framework. She suggests explaining algorithmic decision making in contrast to humans' judgment. This framework is an analogy to Theory of Mind, which regards an understanding of another person's mental states (Logg, 2021). Logg (2021) transferred the framework of Theory of Mind to the 'Theory of Machine' which examines how users perceive differences in human and algorithmic thinking and judging. Besides recognizing how to use AI systems or the plain technical background, the goal is to increase understanding of how the algorithm makes judgments, or in different words, how AI 'thinks' compared on humans. In contrast, AI can be understood as a 'Black Box', when it is explained merely in technical terms, laymen have little insights to AI's internal processes. Logg proposes, comparing algorithmic judgment to humans gives laymen a clear expectations of AI's internal processes. It is important to notice, Theory of Machine implies a change in end-users understanding of the AI's judgment in contrast to humans, rather than altering the AI to become perceived more human-like. For an illustration consider our Theory of Machine briefing in appendix B.

As suggested by Logg (2021), informing individuals about AI processing, implying Theory of Machine, could improve algorithm appreciation. In contrast to most previous interventions to influence algorithm appreciation, establishing better know-how about algorithmic decision making in contrast to human decision making would not affect AI's output, is applicable to all user and is still promising to increase acceptance of algorithmic judgment. To further investigate this suggestion, the general question of this thesis is whether inducing Theory of Machine in a user will increase algorithm appreciation in that user. To our best knowledge, there is no intervention attempting to imply 'Theory of Machine' in an end-user. To test this, we designed a 'Theory of Machine' and Black Box priming to immerse participants into those conditions by influencing their way of thinking about AI (Appendix B). We want to highlight that the goal of our priming is not to present AI as more human-like, as

in anthropomorphism, but to generate a better understanding of AI compared to human judgment.

Of course, the degree to which AI decision are complied with not only depends on the user's conception of that AI system, but also on the difficulty of the task at hand (Hoff, 2015). Gino and Moore (2007) show, that advice seeking is stronger pronounced in hard tasks, while in easy tasks advice is rather rejected. In the context of AI decision support evidence is somewhat mixed. Some studies show people rely more on advice when a task becomes more difficult, as examined by Bogert (2021) in a number guessing task with a human or an AI-based advisor. Here, participants appreciated algorithmic over humans' advice. Moreover, von Walter and colleagues (2021) show AI's advice is more often accepted when the task is merely perceived to be more complex. In their study participants' advice seeking was tested in a real market setting, when developing an interior design concept, and in financial advice. Logg (2019) found no effects of perceived task difficulty on algorithm appreciation across three guessing game tasks on estimating a person's weight, song chart ratings, or romantic attraction. Moreover, her results indicate that participants with expertise in a difficult task tend to reject algorithms advice. Finally, Abeliuk and colleagues (2020) report no effect in a future geopolitical guessing task, in which participants had to assigning a probability to what extend a scenario is possible to happen, for instance the development of long-term interest rate of Canada. Overall, task difficulty seems to be a noteworthy factor when considering AI decision support and, thus far, has only be little investigated. Therefore, we wanted to explore possible moderations of task difficulty between the priming conditions and algorithm appreciation. We consider moderating effects of task difficulty by constructing an easy and hard task and counterbalancing those conditions within the trials. Including this factor allows acquiring further insights on how to establish an approach to increase algorithm appreciation and so fully realize the technology.

The aim of the present study is to investigate the effect of Theory of Machine priming entailing task difficulty on algorithm appreciation. In the experiment, participants either receive a Theory of Machine or black box briefing (Appendix B), which ought to manipulate their perception of the AI advisor. Then, participants perform a guessing game where they display decision making and are supported by an AI advisor. To be concrete, participants need to estimate the age of a person shown in a picture. After their first guess, they receive advise by an AI and have the chance to alter their initial guess. This will be conducted in a total of twenty trials, with ten easy and ten hard to judge pictures. We expect to find that participants in the Theory of Machine condition show more algorithm appreciation than the control group.

Methods

Participants

To conduct the minimum number of participants, we performed a power analysis with the software program G*Power beforehand. Our goal was to obtain .8 power to detect a medium effect size of .5 at the standard .05 alpha error probability. Therefore, a sample size of at least 128 participants is required. We reached the target as 128 participants contributed to the present study, 60 women and 68 men. The mean age was 27 years, ranging from 18 to 66.

Before the initial study, we conducted a pilot to better estimate the length of the study and check for technical issues. The pilot was anonymously distributed to five friends via social media. Afterward those results were deleted.

For the central study, participants were approached on the basis of a data collection platform, Prolific. Participation was compensated with an estimated hourly rate of £7.50, which fluctuates depending on the median completion time. Ultimately, participants received an average of £3 which is corresponding to a workload of around 15 minutes.

Participants were pre-screened to which extent the participant's data was prior approved for other studies on Prolific by an approval rate of 95%, to diminish dropouts. Also, participants were required to be fluent in English, as the manipulation briefing entails complex language.

Participation was voluntary. To ensure an ethical procedure, the study was registered and approved by the Utrecht University Student Ethics Review & Registration Site (UU-SER). Participants signed an informed consent and were debriefed (Appendix C). To ensure anonymity no IP information or demographic data which might reveal the participant's person was collected. Participants were enumerated and data was only linked to their participant number.

Materials

Based on Logg's (2021) and Bogerts and colleagues (2021) experiments, we used a visual estimation task in which participants had to estimate the perceived age of faces presented as pictures by typing their guess into a text box below the image (for an example of the display see Figure 1). After each response, participants were advised by a fictional AI system. After each guess, the AI's suggestion was shown and the participant had a chance to revise their first guess.

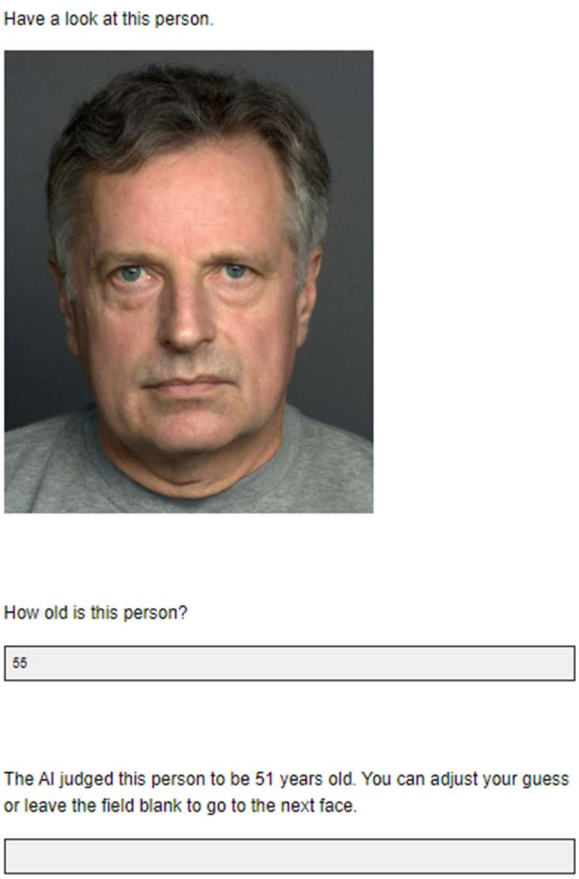

Figure 1: Display of age guessing task, after logging in first age estimate

The faces were retrieved from the FACES database (Ebner et al., 2010). In order to manipulate task difficulty, we considered the standard deviation of the perceived age scores from the FACES database. A high standard deviation implies individuals ranking those faces are inconsistent on the age of the present picture. Thus, it is challenging to estimate the real age of this person. Ten faces were randomly chosen with the smallest standard deviation of the perceived age scale (mean SD = 4.2) for the easy condition, and ten faces were randomly chosen with the highest standard deviation of the perceived age scale (mean SD = 8) for the hard condition. The faces were presented in separated blocks that were counterbalanced.

Half of the participants were assigned to the Theory of Machine condition and the other half to the control condition (Appendix B). In the control condition, the algorithm is presented as a Black Box, as only a technical description of how AI works as an algorithm is given. In the Theory of Machine condition, the participants gain an understanding of the algorithm as in Theory of Machine. This was done by giving instructions on how to use the AI as well as explaining how AI and human decision-making differ.

The main dependent variable was the Weight on Advice (see also Logg, 2019; Bogerts et al., 2021). This measurement computes the extent to which participants align their answers to the algorithm's advice. Here, the difference between the initial and revised judgment divided by the difference between the initial judgment and the advice was calculated. A Weight on Advice score of 0% occurs when a participant ignores advice and a Weight on Advice score of 100% occurs when a participant abandons their prior judgment to match the advice.

Procedure

The entire study was conducted online using the survey provider Qualtrics software XM (www.qualtrics.com). Participants executed the entire procedure online on their own devices by receiving a link via Prolific to the present study. A consent and information letter containing all necessary information was provided in the beginning. Two open-ended questions on demographic data were asked on age and gender. Depending on the randomly assigned condition, participants red the Theory of Machine or Black Box briefing (appendix B). The briefing entailed an awareness check, participants were requested to describe each paragraph of the briefing very briefly in their own words. This was followed by the Visual Estimation Task, where participants estimated ages, and Judge Advisor System, where participants had a change to alter their initial guess after received algorithmic advice.

Task difficulty was included in two different blocks, where participants performed the aforementioned task with ten easy-to-judge and ten hard-to-judge faces. Participants were informed which task difficulty trial (easy or hard) is occurring with a short briefing. The order of both task difficulty conditions was randomly counterbalanced between participants.

After the experiment, participants were asked how certain they were when performing the entire task on a slider from 0 to 100% (Level of confidence, Logg, 2021, appendix A). Finally, participants were debriefed online at the end of the experiment (Appendix C) and redirected to Prolific. On average, participants spend 14 minutes with the experiment, ranging from three to 55 minutes.

Data analysis

Of the 143 participants, a total of 11 incomplete protocols were deleted as well as data of 3 participants who stated not being debriefed after the experiment. At the end of the experiment, participants had the chance to indicate that their data should not be included, in this manner data of one participant was excluded. Moreover, guessing responses were monitored to check if the age ratings were meaningful. We intended to remove participants whose responses deviates more than 3 standard deviations from the group mean in several

trials. No such outliers were detected. Due to technical issues, the level of confidence was recorded in about half of the participants (52%) and is thus not considered in the paper as intended.

With the aid of IBM SPSS 28, the research question was tested, using a 2 (Priming: Theory of Machine/ black box) by 2 (Task Difficulty: hard/ easy) mixed ANOVA with repeated measures on the last factor.

Results

The main effect of priming was not significant, F (1, 126) = .950, p = .332, η^2 = .007. The mean Weight on Advice score was .203 (SD= .204) in the Theory of Machine condition and .174 (SD= .175) in the Control condition. The main effect of Task Difficulty was also not significant, F (1, 126) = 1.653, p = .201, η^2 = .013. The mean Weight on Advice score was .199 (SD=185) in the Easy condition and .177 (SD=196) in the Hard condition. The interaction between Priming and Task Difficulty was also not significant, F (1, 126) = 1.124, p = .291, η^2 = .009).

In addition to the main analysis, we conducted exploratory analyses because we noticed overall low Weight on Advice scores (M = .19). We considered participants' tendency to change, that is, the extent to which they adjusted their guess to the AI's suggestion. Participants' mean Weight on Advice score was assessed. Forty four participants corrected their guesses to less than 10%, 78 participants between 10% and 50%, and six participants by more than 50%. This indicates an overall low tendency to change. We performed a second repeated measures ANOVA as described in the main analysis using only the data from participants whose tendency to change was larger than 10%. Due to the exclusion, data of 84 participants were considered, 41 in the Control condition and 45 in the Theory of Machine condition. The follow-up analysis yields similar results as the main analysis and so revealed no further insights into our data. The main effect of priming was not significant, F (1, 84) = .369, p = .545, η^2 = .004. The mean Weight on Advice score was .274 (SD= .397) in the Theory of Machine condition and .254 (SD = .169) in the Control condition. The main effect of Task Difficulty was also not significant, F (1, 84) = .728, p = .396, η^2 = .009. The mean Weight on Advice score was .275 (SD = .176) in the Easy condition and .254 (SD = .194) in the Hard condition. The interaction between Priming and Task Difficulty was also not significant, F (1, 84) = .757, p = .387, η^2 = .009).

Next, we checked if the task difficulty manipulation was successful. Therefore, we calculated the variance between participants' first guess and the perceived age score from the

FACES database per trial. The mean of all trials per participant was used for the analysis. We conducted a t-test for independent samples. There was not a significant difference in the scores for the task difficulty between the Easy (M = 2.814, SD = .929) and Hard condition (M = 3.227, SD = 509), t (18) = 1.235, p = .233. These results suggest that the task difficulty manipulation was not successful.

Finally, we ran a second analysis on whether the participants' responses were meaningful. In contrast to our exclusion criterion, we analyzed how many initial age estimates were 2.5 standard deviations outside the perceived age score of the FACES Database. The results of 5 participants were always within the interval. Ninety-nine participants' scores were outside the interval between one to four times out of 20 trials, 21 participants scored five to nine times outside the standard deviation, and one participant 14 times. Again, we performed a repeated-measures ANOVA as described in the main analysis using only data of trials falling within 2.5 standard deviations. We thus excluded trails outside this criterion so that the mean Weight on Advice score was calculated with less than the total 20 trails for 123 participants. Across participants in total 482 trails out of 2.560 were removed, meaning about 81% of the initial data was used. This investigation yielded comparable results to the main analysis and so revealed no further insights into our data. The main effect of priming was not significant, F (1, 126) = 1.114, p = .293, η^2 = .009. The mean Weight on Advice score was .182 (SD= .191) in the Theory of Machine condition and .152 (SD = .182) in the Control condition. The main effect of Task Difficulty was also not significant, F (1, 126) = .337, p = .562, η^2 = .003. The mean Weight on Advice score was .171 (SD = .186) in the Easy condition and .163 (SD = .188) in the Hard condition. The interaction between Priming and Task Difficulty was also not significant, F (1, 126) = 1.965, p = .163, η^2 = .015).

Discussion

AI decisions tend to be more accurate than humans, yet this technology can only be fully realized when end-users collaborate with AI. The purpose of this study was to gain a better understanding of how to increase end-users' compliance in algorithm's advice for decision-making tasks. In particular, we examined the effect of a Theory of Machine as opposed to a Black Box priming on algorithm appreciation, when taking task difficulty into account. In the Theory of Machine condition, participants were instructed about the differences between human and AI thinking and judging. In the Black Box condition, only information on how the AI works as an algorithm was presented. We hypothesized that participants in the Theory of Machine condition show more algorithm appreciation than the

control group. This hypothesis was based on previous work on overcoming algorithm aversion by explaining algorithmic judgment to end-users (Cadario et al., 2021; Burton et al., 2019; Logg 2019; Yeomans et al., 2019; Goodwin et al., 2013). We also investigated possible moderations of task difficulty. More specifically, we controlled for interaction between task difficulty and Theory of Machine priming on algorithm appreciation. In contrast to our hypotheses, the results show no main effect of type of priming or a moderation of task difficulty on algorithm appreciation. Theory of Machine and Black Box priming, as well as high and low task difficulty, did not differ significantly from each other in terms of participants' mean Weight on Advice score. Likewise, our results suggest that task difficulty does not influence algorithm appreciation. Furthermore, additional exploratory analyses were performed on participants' tendency to change, tests on whether participants' age ratings were meaningful as well as a task difficulty manipulation check.

This finding may question the relation between inducing a Theory of Machine mindset in an end-user and their behaviour of compiling to algorithmic advice, or more broadly the relationship between mindset and behavioral change. A direct link between mindset change interventions leading to behaviour change was assumed. Kurt Lewin (1946) theorised in his behaviour equation model, that individuals' behaviour results from their personality and environment. This puts into question, if through a single mindset change intervention individuals are going to alter their behaviour. Also, the Theory of Planned Behaviour is prevalent in explaining how behaviour is formed and lays the foundation for decision-making (Ajzen, 1991). According to the theory, individuals display behaviour depending on their attitudes, subjective norms, and perceived behavioral control. Besides predicting behaviour with the Theory of Planned Behaviour, it is also possible to design behaviour change interventions on its basis. Ajzen and Schmidt (2020) found in their meta-analysis, that interventions targeting changes in attitude indeed change attitudes, yet have a small influence on change intentions or behaviour. Interventions should be designed to influence behaviour in the first place, rather than first targeting beliefs or attitudes. In terms of future research, it would be useful to extend the current findings by examining how to realise this approach for to increase behaviour linked to algorithm appreciation.

Moreover, the generalisability of our studies measures needs closer inspection. We used the common measure of Weight on Advice, also used in other studies on algorithmic appreciation (Logg, 2019; Bogerts et al., 2021). Weight on Advice indicates to what extend a participant alters their initial estimation to the suggestion they received from another agent, in this case the AI decision aid. If Weight of Advice can be generalised to algorithmic

appreciation can be disputed. The Weight of Advice measure focuses on a task specific behaviour, while algorithmic appreciation implies a general attitude towards AI. Moreover, if algorithmic appreciation can be linked directly to trust in AI calls for closer inspection. It can be argued that algorithmic appreciation can be understood as a positive attitude towards AI, while trust in AI entails a firm belief. In previous literature all three terms/ concepts, Weight of Advice, algorithmic appreciation and trust in AI are usually directly linked. Hence, the extent to which we can apply the findings of our study to the bigger picture, as in trust in AI, thus need to be interpreted with care.

The present results do not support Logg's (2021) Theory of Machine framework, which proposed that individuals' expectations of how human and AI judgment differ will affect their reaction to algorithmic advice. An important difference between our study and other work on algorithmic appreciation was, that we compare two kinds of introducing AI to participants rather than comparing whether participants prefer an AI over a human advisor. In their previous studies on Theory of Machine, Logg and colleagues (2019) compare a human versus an algorithmic advisor in all experiments. Also in other work (e.g. Abeliuk et al., 2021; Bogert, 2021; Cadario et al., 2021; Burton et al., 2019; Yeomans et al., 2019), participants' reactions to humans and algorithmic advisors are compared, yet, to the best of our knowledge, no studies compare presenting AI in a certain framework compared to another AI's presentation. In a similar manner, in medical science it is common to compare a drug treatment with a placebo drug, rather than comparing a drug treatment with no treatment to investigate the effectiveness of the initial drug. Following this line of argumentation, it seems reasonable to compare an AI with another AI agent rather than an AI with a human advisor. In our experiment we compared two frameworks of introducing AI's decision support aids (as in Theory of Machine or lack box), yet our study yields no significant result between the two AI frameworks.

To fully investigate the Theory of Machine framework, its effect needs to be compared to other frameworks presenting how algorithms function. Another possible framework to introduce AI is by explaining how AI works on a technical level by providing explanations of an AI's decision (Mahmud et al., 2022; Cadario et al., 2021; Schmidt et al., 2020). This stream of literature also shows increased algorithm appreciation in participants. Future research on algorithm appreciation needs to investigate if a Theory of Machine framing is more effective than other forms of explaining algorithms to end-users. In other words, future research should challenge several briefings, of introducing AI to an end-users, effects on algorithm appreciation. For other technologies, where collaboration depends on the

end-user, this level of comparison is common. For instance, in social robots, an implicit and explicit mind perception briefing is tested (Keijserset al., 2021), or anthropomorphic to a functional description (Wallkötter et al., 2020; Onnasch & Roesler, 2019). In essence, no clear methodology is currently available to implement Theory of Machine in end-users and future research will be needed to test the framework proposed by Logg. Hence, linking our findings to previous research is difficult. Generally speaking, an increased understanding of AI in contrast to human judgment improves algorithm appreciation theoretically, yet the leaves the question open on if explaining how AI and human decision-making differ is more effective than giving other insights in AI's processing to increase trust in algorithmic decision aids.

At the same, the possibility arises that our manipulation to induce Theory of Machine failed. There are at least three potential limitations concerning the results of this study. First, the validity of the results is limited by alternative explanations. An alternative explanation for our results could be that participants did not see a need for cooperation with the AI decision aid, rather than mistrusting the AI. In other words, participants might have strong opinions when guessing the age of the presented pictures as their tendency to change towards the AI's suggestions was low. A Weight on Advice score of 100% would mean participants fully adjust their guess to the AI's suggestion. In our study, the overall mean Weight on Advice score was 19%. Only minor differences were found between the Theory of Machine (Weight on Advice = 20%) and the control condition (Weight on Advice = 17%). Using Weight on Advice as an indicator for trust in AI was used in other studies and seems like a plausible measurement. Yet, in other studies a single measurement was completed and the experiment was performed in a lab rather than online as in the present study (Mahmud et al., 2022; Bogerts et al., 2021; Logg, 2019). There is no indication that our participants rushed through the survey, as participants passed the attention check questions. Still, repeating the study in a lab and ensuring the administration of the task is performed thoroughly could lead to better considerable results.

Second, another alternative explanation might be that our control condition had the same effect as our experimental condition, namely increasing algorithm appreciation. In our control condition, we described algorithms as a black box by only giving technical details on how AI operates. Both of our conditions might have led to a better general understanding of AI. Cadario and colleagues (2021) found that participants reporting a high understanding of AI decision aids increases trust in that participant. In their study, they measured participants' reported subjective understanding of medical AI or a human doctor in skin cancer diagnosis.

Individuals who indicated a high subjective understanding of the algorithm were more likely to use AI as a service to detect skin cancer. In their second study, an intervention was conducted using Google ads on AI-based skin cancer detection. The ad stated how the AI processes or does not process data. People were more likely to respond to the ads where an explanation was given. All in all, their work suggests that participants preferred algorithms when they indicate a subjective understanding of the AI. This might show how any form of increasing participants' understanding of AI can lead to algorithm appreciation, which in turn could mean our control condition, explaining AI as an algorithm, increased trust in AI. For future research, we propose to replicate the Theory of Machine framing but modify the control group to anticipate the study set up to become effective.

Third, another possible source of error is that our priming intervention might have failed to induce a Theory of Machine mindset in participants. Some recent criticisms of priming interventions are summarised in Weingarten and colleagues (2016) meta-analysis on social-priming literature. The authors expressed doubts about priming due to replication problems of several priming studies and general low effect sizes across methodological procedures. Especially priming through merely introducing a stimulus is little effective. In our study, we attempted to go beyond a passive reading exercise to prime a Theory of Machine or Black Box mindset. Participants had to repeat instructions in their own words so that they encode information they read in the briefing. This exercise could be performed in a more engaging manner. One potential source of inspiration for a more effective way to induce a Theory of Machine mindset may be through generating self-instructions. Generated information is better remembered than when just reading it, also known as the generation effect (Bertsch et al., 2007). Consequently, it can be argued that self-instructions are more powerful than instruction reading interventions. A similar setup is detailed in Cadario and colleagues (2021) studies mentioned earlier. In their procedure, participants had to write down their assumptions about how they expect algorithms to work. Generating own assumptions creates an active way to induce a certain mindset. Yet this way it is harder to manipulate the briefing. More research is needed to determine how to induce a Theory of Machine mindset.

Task difficulty

In the present study, including on task difficulty, no difference between the high and low condition could be found. This may be an indicator that task difficulty is not critically affecting algorithm appreciation. On the one hand, this result is consistent with Abedliuk and colleagues (2020) and Logg and colleagues (2019), who also observed no effect. In Abedliuk and colleagues (2020) work, several aspects are considered besides task difficulty which have

a stronger influence on algorithm appreciation, such as beliefs about AI or cognitive bias (Abeliuk, 2020). One could argue, with a Theory of Machine briefing individuals' perception of AI change. A Theory of Machine manipulation may influence their pre-existing beliefs about AI as well as cognitive biases. By influencing these aspects, we can focus better on the moderating effects of task difficulty. In other words, by reducing the other moderators with a Theory of Machine briefing, namely uncovering beliefs about AI and diminishing cognitive biases, a better focus on the influence of task difficulty on algorithm appreciation can be drawn. Yet since our study's main effect of a Theory of Machine priming is missing, those suggestions about moderating effects of task difficulty need to be interpreted with caution.

On the other hand, work on the potential effects of task difficulty on algorithm appreciation has been conducted. Throughout several decision-making tasks, as described earlier, Bogert and colleagues (2021) and von Walter and colleagues (2021) show how task difficulty can impact compliance with AI's advice. The current study does not support these research findings. However, the shortcoming of our methods should be recognized, as our task difficulty manipulation failed to show a significant difference between the easy and hard task's error rate of participants. Our approach was based on high and low standard deviations of pre-registered age ratings from the FACES database. Conversely, Bogert and colleagues (2021) used a straightforward task difficulty manipulation, where pictures either showed a crowd of 15 or 5000 humans. Participants needed to estimate how many people are shown in a picture. Comparing both tasks, it is more intuitive to consider the difficulty of a picture showing a small or large crowd than predicting the difficulty of assessing an individual's age. The difficulty of age ratings can merely be judged based on rather abstract statistical analysis. Future work should concentrate on enhancing the quality of task difficulty manipulations. Then further experimental investigations can be realized to estimate the effects of task difficulty on algorithm appreciation.

Conclusion

Algorithm-based decision support systems are common in our day-to-day life. However, how to encourage individuals to appreciate AI's advice is open to debate. Perceptions of algorithms are a novel field in psychology where theoretical approaches need to be further established to enable practical implications to fully realize the technology.

Literature indicates how explaining AI not simply on a technical level, but rather in comparison to human cognitive processes might lead to more algorithm appreciation. The present study represents a first attempt to translate the theoretical framework of Theory of Machine into practice. Despite we found no effect of our Theory of Machine framing on

algorithm appreciation, we believe our work could be a starting point to establish a method to increase trust in AI. Considering strong support of literature, we believe the concept of explaining how AI 'thinks' promising to improve algorithm appreciation (Logg, 2021; Burrton, 2019; Yeomans et al., 2019). In contrast to other practical implications, focusing on improving algorithms technically or letting the end-user influence the algorithm's judgment or outcome, priming end-users with Theory of Machine might solve the issue without altering the technology. Further research is needed to determine whether a Theory of Machine based intervention can influence end-users to act upon recommendations posed by algorithms to deal with the last mile problem.

References

Abeliuk, A., Benjamin, D. M., Morstatter, F., & Galstyan, A. (2020). Quantifying machine influence over human forecasters. *Scientific reports*, *10*(1), 1-14.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Ajzen, I., Fishbein, M., Lohmann, S., & Albarracín, D. (2018). The influence of attitudes on behavior. *The handbook of attitudes*, 197-255.

Ajzen, I., & Schmidt, P. (2020). Changing behavior using the theory of planned behavior. *The handbook of behavior change*, 17-31.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & cognition*, *35*(2), 201-210.

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, *11*(1), 1-9. *Humans rely more on algorithms than social influence as a task becomes more difficult*

Bonnefon, J. F., & Rahwan, I. (2020). Machine Thinking, Fast and Slow. *Trends in Cognitive Sciences*, *24*(12), 1019–1027.

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.

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12), 1636–1642.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

de Fine Licht, Karls., & de Fine Licht, Jenny. (2020). Artificial intelligence, transparency, and public decision-making. *AI & society*, *35*(4), 917-926.

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.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79-94.

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES - A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods, 42*(1), 351-362. doi:10.3758/BRM.42.1.351.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of personality assessment*, *50*(3), 387-395.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*(6), 570-576.

Fishbein, M., & Ajzen, I. (2011). *Predicting and changing behavior: The reasoned action approach*. Psychology press.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*.

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

Goodwin, P., Gönül, M. S., & Önkal, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, *29*(2), 354-366.

Goodyear, K., Parasuraman, R., Chernyak, S., de Visser, E., Madhavan, P., Deshpande, G., & Krueger, F. (2017). An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social neuroscience*, *12*(5), 570-581.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of applied psychology*, *98*(6), 1060.

Lewin, K. (1946). Behavior and development as a function of the total situation. In L. Carmichael (Ed.), *Manual of child psychology* (pp. 791–844). John Wiley & Sons Inc

Logg, J. (2021). The Psychology of Big Data: Developing a "Theory of Machine" to Examine Perceptions of Algorithms. *Peer Reviewed and Accepted in Matz, S.(Ed.), American Psychological Association Handbook of Psychology of Technology*.

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.

Mahmud, H., Islam, A. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57-87.

Molden, D. C. (Ed.). (2014). *Understanding priming effects in social psychology*. Guilford Publications.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260-278.

von Walter, B., Kremmel, D., & Jäger, B. (2021). The impact of lay beliefs about AI on adoption of algorithmic advice. *Marketing Letters*, 1-13.

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, *364*.

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological bulletin*, *142*(5), 472.

Westin, C., Borst, C., & Hilburn, B. (2015). Strategic conformance: Overcoming acceptance issues of decision aiding automation?. *IEEE Transactions on Human-Machine Systems*, *46*(1), 41-52.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403-414.

Appendix A

Question: Level_of_Confidence

How likely is it that your estimates were the persons' actual age?

Type a number to indicate your level of confidence (0 means no chance, 100 means absolutely certain).

Appendix B

Theory of Machine and Black Box Condition Briefing

Description	**Theory of Machine Condition**	**Black Box Condition**
Intro	In this task you will see randomly chosen pictures of different persons and guess their age. It is important that you do a precise guess because this will influence the learning of the AI. After you submit your result, the AI's result will be presented. To validate the AI, you can adjust your guess if needed. Before we start with the task, you will receive a quick introduction on how to work with the AI. Also, we want to inform you about the differences between human and AI thinking and judging.	In this task you will see randomly chosen pictures of different persons and guess their age. It is important that you do a precise guess because this will influence the learning of the AI. After you submit your result, the AI's result will be presented. To validate the AI, you can adjust your guess if needed. Before we start with the task, you will receive a quick introduction on how to work with the AI. Also, we want to inform you how the AI works as an algorithm.
Input (the information used)	The information used by the algorithm stems from a face database. A lot of data is needed: the larger the dataset is the better the algorithm can find patterns. Most algorithms require an element of human judgment, whether to determine the input data or to interpret the output. Therefore, the algorithm is only as good as input from humans provided.	The information used by the algorithm stems from a face database. A lot of data is needed: the larger the dataset is the better the algorithm can find patterns. Most algorithms are based on artificial networks. These networks consist of nodes that are connected to each other. When information, such as a picture is presented, it is decomposed by activating specific

	In contrast to the way humans use data, the AI processes a huge amount of data input. Humans cannot capture those greater quantities of data. In your own words, can you **!very briefly!** describe how AI uses data and what the difference is to humans?	nodes, which in turn activate connected nodes. This way the information is passed on through the network and analysed. In your own words, can you **!very briefly!** describe how AI uses data?
Process (how the same information is utilized)	On this basis the AI processes information and calculates results. The AI considers facial characteristic, finds patterns and links those to an age. The AI 'trains' for a specific task, namely age recognition based on faces. The AI is able to improve on its own, with every trial the AI gets more precise in guessing the right age. Just as in human judgment, AI's judgment is not perfect. Yet, the AI makes decisions with higher accuracy compared to humans. Human analysts have limited time and brainpower to process and analyse this data. Moreover, the AI is able to process data perfectly rational. What makes	On this basis the AI processes information and calculates results. The AI considers facial characteristic, finds patterns and links those to an age. The AI 'trains' for a specific task, namely age recognition based on faces. The AI is able to improve on its own, with every trial the AI gets more precise in guessing the right age. The training procedure calls upon the backpropagation algorithm. The connections between nodes in a network have different weights. These weights are adapted during training and such that nodes become connected in specific way. Information in a network is stored in a distributed way. This means

	it more objective than human analysis is that AI excludes individual errors, biases and preferences as it unites a huge amount of data from various people. In your own words, can you **!very briefly!** describe how AI processes data and what the difference is to humans?	that one node only represents a tiny bit of information and that complex information such as a face and its corresponding age need a tremendous number of nodes that are all connected to each other with specific weights. In your own words, can you **!very briefly!** describe how AI processes data?
Output (the predictions, advice, and feedback that are produced)	Finally, the AI informs you about its result. The AI just presents its predicted age, without giving you additional context information to its result. When humans provide an answer, they tend to personalise it or add an interpretation to it by the way they describe their response. In your own words, can you **!very briefly!** describe how AI gives you feedback and what the difference is to the way you would receive feedback from humans?	Finally, the AI informs you about its result. The previous calculations produce a result based on the chosen output type. In this case it is a number representing human age in years. In your own words, can you **!very briefly!** describe how AI gives you feedback?
Outro	Thank you for completing the introduction! Now we can start with the task.	Thank you for completing the introduction! Now we can start with the task.

Appendix C

Debriefing

A fictional framework was created to increase the validity and authenticity of the task. The algorithm mentioned throughout the task is fictional. Hence, the participants believe an AI advisor is suggesting the person's age while those answers were predesigned by us. The participants were disclosed about the incorrect information and why this was necessary for the experiment in the debriefing.

Debriefing as received by the participants:

Thank you for your participation in this research study. For this study, it was important that I provide you with incorrect information about some aspects of the study. Now that your participation is completed, I will describe the incorrect information to you, why it was important, answer any of your questions, and provide you with the opportunity to make a decision on whether you would like to have your data included in this study.

What you should know about this study

You were informed, that the aim of this research is to study the accuracy of an AI based decision support system. Whereas the actual aim of the study is to investigate how to increase algorithm appreciation in AI based decision support systems, in other words, to what extend to people accept the advice from AI. Hence, training the AI system was not the point of attention but your adjustments after seeing the AI's advice. Also, the AI system does not exist. Those answers were randomly predesigned by the researcher. This was necessary for the study to measure your true reactions towards a potential AI support system.

If you have questions

The main researcher conducting this study is Katharina Olejnikov, a master's student at the Utrecht University. Please ask any questions you have by contacting me via k.olejnikov@students.uu.nl. If you have an official complaint about the investigation, you can send an e-mail to the complaints officer via klachtenfunctionaris-fetcsocwet@uu.nl.