Slow response times undermine trust in algorithmic (but not human) predictions

Organizational Behavior and Human Decision Processes 157 (2020) 103–114

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Slow response times undermine trust in algorithmic (but not human) predictions☆

Emir Efendić[a,*], Philippe P.F.M. Van de Calseyde[a], Anthony M. Evans[b]

[a] Eindhoven University of Technology, the Netherlands
[b] University of Tilburg, the Netherlands

ARTICLE INFO

Keywords:
Response time
Judgment and decision making
Prediction
Algorithm aversion
Human-computer interaction

ABSTRACT

Algorithms consistently perform well on various prediction tasks, but people often mistrust their advice. Here, we demonstrate one component that affects people's trust in algorithmic predictions: response time. In seven studies (total $N = 1928$ with 14,184 observations), we find that people judge slowly generated predictions from algorithms as less accurate and they are less willing to rely on them. This effect reverses for human predictions, where slowly generated predictions are judged to be more accurate. In explaining this asymmetry, we find that slower response times signal the exertion of effort for both humans and algorithms. However, the relationship between perceived effort and prediction quality differs for humans and algorithms. For humans, prediction tasks are seen as difficult and observing effort is therefore positively correlated with the perceived quality of predictions. For algorithms, however, prediction tasks are seen as easy and effort is therefore uncorrelated to the quality of algorithmic predictions. These results underscore the complex processes and dynamics underlying people's trust in algorithmic (and human) predictions and the cues that people use to evaluate their quality.

1. Introduction

Individuals and organizations increasingly rely on algorithmic predictions.[1] Such interactions, where a person receives advice generated by an algorithm and decides on its implementation, constitute a crucial part of modern workflows (Willson, 2017). For example, algorithmic predictions are an everyday feature in many organizations to aid in sales forecasting (Fildes & Goodwin, 2007), in medical situations (Stacey et al., 2017), and even in matters related to justice (Porter, 2018). To boot, algorithms often outperform humans, producing predictions of superior quality (Beck et al., 2011; Carroll, Wiener, Coates, & Galegher, 1982; Dawes, 1971; Meehl, 1954; Youyou, Kosinski, & Stillwell, 2015) although there have been instances where they have produced biased advice (O'Neil, 2016; Wachter-Boettcher, 2017). However, repeated observations show that people profoundly mistrust algorithm-generated advice, especially after seeing the algorithm fail (Bigman & Gray, 2018; Diab, Pui, Yankelevich, & Highhouse, 2011; Dietvorst, Simmons, & Massey, 2015; Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009).

What affects people's trust in algorithmic predictions? The present research addresses this question by investigating a common feature in the prediction process. More specifically, we propose that the speed with which a prediction is generated affects people's trust in algorithmic predictions. Just like with human forecasters, algorithms can take varying degrees of time to generate predictions – a feature that can become highly salient when a user interacts with the same algorithm over a long period of time. In various industries, forecasters use the same algorithmic support system to make predictions about future sales, orders, or hiring decisions (Power, 2002). What are the consequences of observing variations in an algorithm's prediction speed? Are people more likely to trust predictions that an algorithm generated almost immediately or after a long pause? We report seven studies that systematically test how the speed with which algorithms generate predictions (fast versus slow) impacts people's willingness to trust these predictions. We contrast this with how the prediction speed of others affect an observer's willingness to trust their prediction. This provides us with insights into how the same cue (i.e., response time) can be interpreted differently as a function of different prediction providers

☆ This work was supported by the TKI Dinalog funding agency on the project: "Increasing the usability, acceptance, and adoption of advanced planning and scheduling systems"; Grant n.: 2016-1-074TKI.

* Corresponding author at: School of Business and Economics, Maastricht University, 6211 LM Maastricht, the Netherlands.
E-mail address: e.efendic@maastrichtuniversity.nl (E. Efendić).

[1] For the purposes of this paper, we loosely define "algorithm" to include any evidence-based forecasting formulas and rules such as statistical models, decision aids, or other mechanical procedures (Dietvorst et al., 2015).

https://doi.org/10.1016/j.obhdp.2020.01.008
Received 7 April 2019; Received in revised form 19 January 2020; Accepted 21 January 2020
Available online 07 February 2020

(i.e., algorithmic- vs. human-generated predictions).

The article is organized as follows. We start by examining the recent literature in psychology and economics on how people interpret human response times in social interactions. We subsequently discuss how different response times may influence trust in algorithmic predictions. We then describe our experimental tests and conclude with a broader discussion of the results.

2. Prediction accuracy and response time as information

In recent years, researchers in psychology and economics have looked at how observing others' response times influences various interpersonal judgments and behaviors (Critcher, Inbar, & Pizarro, 2013; Evans & van de Calseyde, 2017; Konovalov & Krajbich, 2017; Mata & Almeida, 2014; Van de Calseyde, Keren, & Zeelenberg, 2014). For decisions based on preferences, people believe that others' response times are associated with feelings of doubt or conflict. For example, Critcher et al. (2013) asked participants to evaluate the moral character of two persons who found wallets filled with cash. Both decided to keep the wallet, but one made the decision relatively quickly, whereas the other made the same decision slowly. In turn, the person who was slower to decide to keep the wallet was judged as less dishonest than the one who immediately chose to keep it (see Van de Calseyde et al., 2014 for how others' response times affect interpersonal choices).

In explaining these effects, the above-mentioned research found that people use observed response times as information. That is, slow decisions signaled feelings of conflict and doubt to observers (whereas fast decisions signaled confidence), explaining why people evaluated the person who was relatively slow in choosing to keep the wallet as less dishonest. However, slow response times are perceived differently for tasks that people presume require *effort* (e.g., making difficult predictions). In such cases, observing slower response times indicates that the person exerted the necessary effort to complete the task, whereas faster responses reveal a lack of effort or commitment (Jago & Laurin, 2018; Kupor, Tormala, Norton, & Rucker, 2014). Importantly, the more effort people believe others invest in completing relatively difficult tasks, whether in the form of time, physical exertion, pain, or money, the more positive the outcome of that effort is evaluated (Festinger, 1957; Kruger, Wirtz, Van Boven, & Altermatt, 2004; Labroo & Kim, 2009; Norton, Mochon, & Ariely, 2012).

In testing this 'effort heuristic', Kruger et al. (2004) asked participants to evaluate the quality of two paintings made by the same artist. In one condition, participants were told that the artist finished the first painting in 18 h, whereas it took her 4 h to finish the second painting. In the second condition, this information was reversed (i.e., 4 h to finish the first, 18 h to finish the second painting). Consistent with the conjecture that people use time spent on completing a task as a heuristic for quality, paintings that took longer to finish were judged as being of higher quality (regardless of the order in which they were made). Here, we argue that the speed with which predictions are generated similarly influences how observers evaluate the quality of predictions. More precisely, given that slow response times and actions lead to perceptions of effort and commitment when completing difficult tasks, observers are expected to perceive others' predictions as being of higher quality when they are generated slowly (versus quickly).

Although slow response times are expected to increase the perceived quality of human-generated predictions, it remains unclear how people would perceive slow *algorithmic* predictions. We propose that people have different expectations of how *difficult* prediction tasks are for algorithms, compared to humans. Some tasks, like image recognition for instance, are extremely easy for humans, but (currently) difficult for algorithms (Krizhevsky, Sutskever, & Hinton, 2012). Conversely, people may think that making a prediction is a relatively easy task for an algorithm, as it is an objective task involving the integration of multiple pieces of information or complex calculations (Castelo, Bos, & Lehmann, 2019). This view leads us to predict that slower response times will lead to lower quality evaluations of algorithmic predictions, as they will signal more effort being exerted for an ostensibly easy task.

This proposition is based on the notion that people perceive the quality of advice differently depending on whether the advice provider has engaged in the right amount of thinking required by the situation (i.e., when their level of thoughtfulness matches the apparent difficulty of the task). For example, Kupor et al. (2014) found that more thoughtful decisions (varied by describing how much effort was devoted) were seen as higher in quality, *but only for difficult decisions*. For easy decisions, the findings were less clear: more thoughtful decisions were generally seen as lower in quality, but the amount of thinking did not always have a statistically significant impact. This work suggests that the relationship between effort and perceived quality thus depends on observers' beliefs about task difficulty.

If people think that prediction tasks are easy for algorithms, then longer responses ought to lead to decreased prediction quality evaluations because the algorithm's level of effort would not match the apparent difficulty of the task.[2] Conversely, one can predict that for tasks which people consider to be difficult for an algorithm, longer response times ought to lead to increased quality evaluations. However, we maintain that people will consider most prediction tasks to be easy for algorithms. Therefore, we should observe that longer response times *generally* lead to lower quality evaluations for algorithmic predictions.

3. Present research

We conducted seven studies (see Table 1 for an overview) to test how people judge the quality of algorithmic- and human predictions. Using a variety of different prediction contexts and methodologies, we find that slow human predictions are judged as superior to fast human predictions. However, the opposite occurs for algorithms: fast algorithmic predictions are judged as superior to slow algorithmic predictions. While speed impacts perceptions of effort similarly for both algorithms and humans (i.e., slower speeds lead to perceptions of more effort being exerted), the relationship between perceived effort and prediction quality differs for humans and algorithms because people perceive prediction tasks to be easy for algorithms, but difficult for humans.

At the same time, we also observe that response time is a more evaluable attribute for humans than for algorithms as it has an impact both in joint (within-subject) and single (between-subject) evaluation conditions. While the effect of response time can appear in single-evaluation conditions for algorithms, this is moderated by the user's previous experience with the algorithm (i.e., slower predictions were judged as increasingly worse over time). Finally, we find that these inferences have behavioral consequences: people are more likely to choose a human-generated prediction over a slowly generated algorithmic prediction. Additionally, in an incentivized study using sports predictions, we find that people are more willing to rely on quick (as opposed to slow) algorithmic predictions.

For all studies, we report how we determined the sample size, all data exclusions (if any), all manipulations, and all measures. All studies but one (Study 5) were pre-registered. The links to the registrations are provided in the appendix, where we also provide a link to the projects' OSF page with access to data, materials, and analysis code.

Data were analyzed using multi-level models with random estimates for participants and varying different prediction scenarios and response times across participants (Westfall, Kenny, & Judd, 2014). We relied on the lme4 (Bates, Mächler, Bolker, & Walker, 2015) and the lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017) packages in R to construct the models and extract p-values. Since there are currently no

[2] More conservatively, the findings from Kupor et al. (2014) suggest that there should be no positive relationship between effort and perceived quality for easy tasks (if not a significant negative relationship).

Table 1
Overview of studies.

Study	Prediction provider	Prediction domain	DV	Main finding
1	Alg & Hum	Student success	Prediction accuracy	Slow alg/Fast hum, less accurate
2	Alg & Hum	Sales	Prediction accuracy	Slow alg/Fast hum, less accurate
3	Alg & Hum	Sales	Prediction accuracy	Slow alg/Fast hum, less accurate; Effort as mediator
4a/b	Alg & Hum	Sales/Employee absence	Prediction accuracy	Difficult, slow predictions more accurate; Easier, fast predictions more accurate
5	Alg	Sports	Prediction accuracy & persuasiveness	Slow alg less accurate; effect stronger with more experience with algorithm
6	Alg	Sales	Consulting a human over an algorithm	More additional consultation for slow alg
7	Alg	Sports	Prediction accuracy & persuasiveness/ choice	Slow alg, less accurate; more fast predictions chosen

widely accepted effect size estimates for multi-level models we report standard Cohen's *dz*.

4. Studies 1 and 2

In Studies 1 and 2 we investigated the impact of fast- versus slow response times on the perceived accuracy of human- vs. algorithmic predictions. We hypothesized that slow human predictions would be evaluated as *more* accurate than fast human predictions, whereas slow algorithmic predictions would be evaluated as *less* accurate than fast algorithmic predictions. Both studies followed a similar procedure so we describe them together.

4.1. Methods

4.1.1. Participants

Both studies were conducted on MTurk. Participants were assigned to a 2 (Prediction provider: Human vs. Algorithm; between-subjects) × 2 (Response time: Fast vs. Slow; within-subjects) mixed-design experiment. After excluding participants who did not pass the initial attention check and those who did not complete the entire study, there were 304 participants (46% female; M_{Age} = 36.45, SD_{Age} = 11.28) in Study 1 and 302 participants (47% female; M_{Age} = 38.79, SD_{Age} = 12.15) in Study 2.

4.1.2. Procedure

The two studies differed in the task scenarios used and whether an actual prediction, ostensibly made by a human or an algorithm, was shown. In Study 1, participants were told to imagine that they were an admissions officer working at a public university where they had to predict the academic success of potential students. They were then told that admission officers receive various pieces of information about each student and that this information is used to make predictions about the student's success. In Study 2, participants were told that they were sales officers working for a large consumer goods company and that their task was to predict the future sales of various products.

Participants were told that because of university (S1) or company (S2) regulations, as a quality assurance measure, one always needs to consult a colleague [an algorithm] when making a prediction. Additionally, they were told that they would know how much time the colleague [algorithm] took to generate the prediction. In Study 2, participants were also told that the company uses "boxes" to represent sales units and that a sales officer might predict future sales of an X number of boxes of a specific product. So, for each product, we provided participants with a prediction of boxes, ostensibly made by a human colleague [algorithm]. The predictions could vary randomly from 10 to 90 boxes, in increments of ten.

Participants went through six randomly presented vignette scenarios, each representing an individual student (S1) or product (S2). Three of the predictions were described as provided quickly and three as provided slowly. The response time descriptions varied. For the fast predictions we used: "*after only a couple of seconds*", "*immediately*", and "*straight away*". For the slow predictions we used: "*after a long pause*", "*after some time*", and "*after an extended period of time*". No additional information about the colleague was provided. In the algorithm condition, the participants were told that the statistical algorithm is called "StatCast" and that it was designed by the university/company to predict the success of students (S1) or future sales (S2).

Participants evaluated what they thought the accuracy of the prediction was on a scale from −3 (*very inaccurate*) to 3 (*very accurate*).[3] In addition, after providing all six of the accuracy estimates, each participant responded to two questions (one for fast and one for slow speeds – presented randomly) on how likely they would have been to use the prediction as their own (−3 *very unlikely* to 3 *very likely*).

4.2. Results[4]

4.2.1. Perceived accuracy

A 2 (human = −0.5; algorithm = +0.5) × 2 (fast = −0.5; slow = +0.5) analysis found a significant effect of the prediction provider in S1, $F(1, 303) = 3.97$, $p = .05$, $dz = 0.11$ and in S2, $F(1, 300) = 23.77$, $p < .001$, $dz = 0.28$. Algorithms were considered more accurate overall, compared to humans. In S1, there was also a main effect of response time, $F(1, 303) = 4.59$, $p = .03$, $dz = 0.12$. Slow predictions were considered as more accurate compared to fast predictions. In S2, there was no main effect of response time ($F < 1$). Most importantly, there was a two-way interaction in both S1, $F(1, 303) = 25.03$, $p < .001$, $dz = -0.29$ and S2, (1, 300) = 13.36, $p < .001$, $dz = -0.21$ (see Fig. 1, Study 1-A and Study 2-C subplot).

Next, we compared the simple effect of response time for human- and algorithmic predictions. Both in S1, $F(1, 156) = 18.82$, $p < .001$, $dz = 0.25$ and in S2, $F(1, 154) = 6.84$, $p = .01$, $dz = 0.15$, participants evaluated the accuracy of human-generated predictions as much higher when it was generated slowly, than when it was generated quickly. Similarly, both in S1, $F(1, 147) = 4.07$, $p = .05$, $dz = -0.11$ and in S2, $F(1, 146) = 5.75$, $p = .02$, $dz = -0.14$, participants evaluated the accuracy of algorithm-generated predictions as much lower when it was generated slowly, than when it was generated quickly.

4.2.2. Willingness to use predictions

Using the same analysis approach as above, we again found significant main effects of the prediction provider in S1, $F(1, 303) = 4.05$, $p = .05$, $dz = 0.12$ and in S2, $F(1, 300) = 4.47$, $p = .04$, $dz = 0.12$. There was again a main effect of response time in S1, $F(1, 303) = 8.85$, $p = .003$, $dz = 0.29$, but not in S2. Both effects were in the same

[3] The scale was re-coded to range from 1 to 7 in the analysis. This was the case in all studies that used these anchors.

[4] In the preregistration we stated that we would perform mixed ANOVA's *and* regressions. We report the regressions to be in line with the other presented studies. However, the data analysis files (https://osf.io/efauv/) contain code for performing the ANOVA's which show the same results.

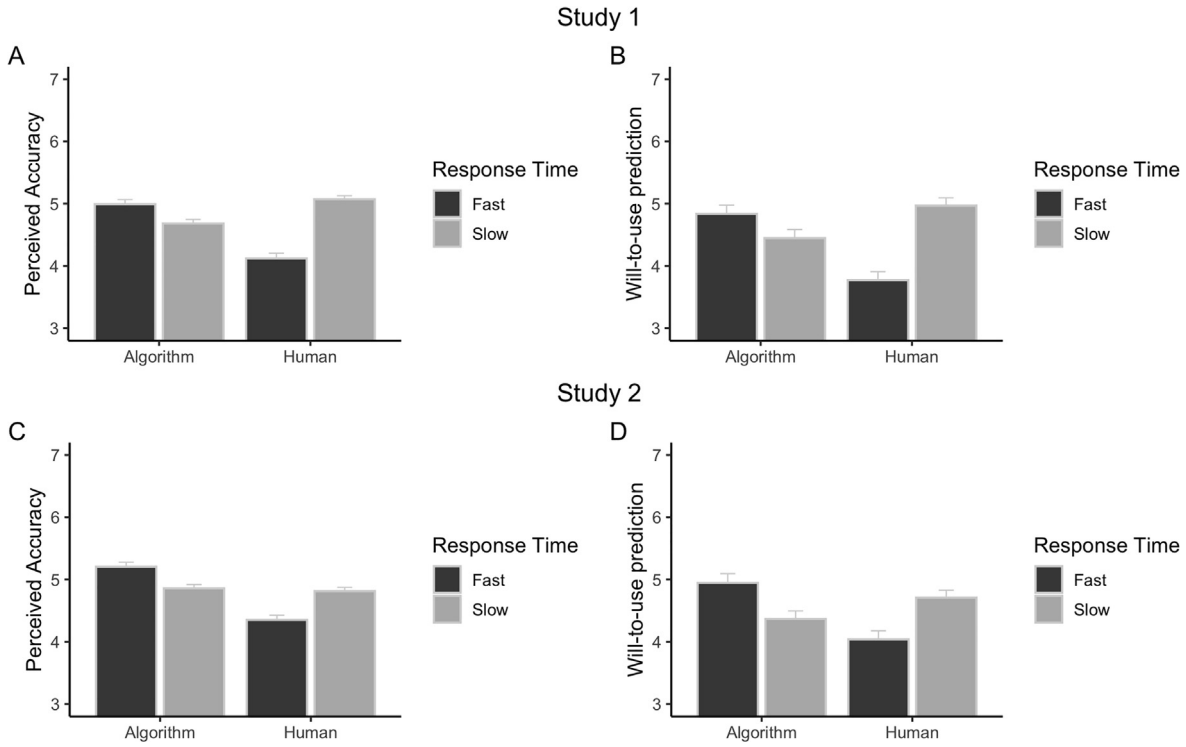

Fig. 1. The means and standard errors of Study 1 (upper row) and Study 2 (lower row) results on perceived accuracy of the generated prediction (A and C) and willingness to use generated prediction (B and D) as a function of prediction provider (Algorithm vs. Human) and response time (Fast vs. Slow).

direction as in the analysis above. Importantly, there was again a significant two-way interaction in both S1. $F(1, 303) = 34.44$, $p < .001$, $dz = -0.57$ and S2, $F(1, 300) = 21.89$, $p < .001$, $dz = -0.27$ (see Fig. 1, Study 1-B and Study 2-D subplots). Simple effects showed that for the human-generated predictions, participants were more willing to use those predictions that the colleague generated slowly in S1, $F(1, 156) = 41.19$, $p < .001$, $dz = 0.37$ and in S2, $F(1, 154) = 13.61$, $p < .001$, $dz = 0.21$. The reverse was true for algorithmic predictions in S1, $F(1, 147) = 4.00$, $p = .05$, $dz = -0.11$ and in S2, $F(1, 146) = 8.71$, $p = .004$, $dz = -0.17$; participants were more likely to use quickly generated algorithmic predictions.

4.3. Discussion

The first two studies demonstrate that the response time cue has differential effects on the perceived accuracy of human- versus algorithmic predictions. Specifically, slowly generated human predictions were seen as more accurate. However, this reversed for algorithms (i.e., slow predictions were seen as less accurate). Importantly, this result also extended to a person's willingness to use a prediction as their own (i.e., a greater willingness to use slowly generated human predictions, but a lower willingness to use slowly generated algorithmic predictions). These effects replicated across two different task scenarios and when participants were provided with actual numeric predictions. Our next study investigates the mechanism underlying the different effects of response time on the perceived quality of human- vs. algorithmic predictions.

5. Study 3

The first two studies demonstrated that the relationship between response time and prediction quality differs for human- vs. algorithmic predictions. Building on these results, we test a moderated mediation model where slower response times are seen as signaling more effort for both algorithms and humans. However, we predict that the relationship between perceived effort and prediction quality is moderated by the prediction provider. This moderation is related to differences in the perceived difficulty in making predictions for humans vs. algorithms. For human predictions, we expected that the prediction task should be seen as difficult; therefore, more effort should lead to higher quality evaluations (Kupor et al., 2014). For algorithms, the prediction task should be seen as easy. Therefore, more algorithmic effort should not be related to prediction quality, or more effort should lead to lower quality evaluations. To test this account, we conducted a study measuring perceived task difficulty for algorithms/humans, perceived effort, and prediction accuracy.

5.1. Method

5.1.1. Participants

Five hundred and four participants were recruited on MTurk. The study had the same design as Studies 1 and 2. We aimed to recruit 230 people per between-subject condition. After excluding people who failed the attention check or simply did not complete the full study, we had 486 participants (58% female; $M_{Age} = 38.39$, $SD_{Age} = 11.04$) in Study 3.

5.1.2. Procedure

The procedure was similar to Study 2 with three changes. First, we inserted a question asking people how difficult they thought making predictions was for humans/algorithms: "Fill in the blank: Predicting future sales is a task that is relatively ___ for a human [algorithm] to accomplish." Participants could either select "easy" or "difficult". We randomly varied whether this question was presented before or after participants were presented with any of the predictions. Second, after being presented with the speed of the prediction provider, participants were asked: "How much effort did your colleague [StatCast] exert to come to this prediction?". They could answer on a 1 (*Little effort*) to 7 (*Much effort*) scale. Third, because the accuracy question was on a separate screen and after the effort question, we wanted to make sure that

the participants were aware of the response time manipulation. We thus re-worded the question[5] to: "Given your colleague's [algorithm's] delayed [quick] response time, how accurate do you think is his [its] prediction?"

5.2. Results

As expected, most people (81.07%) thought making predictions is a difficult task for a human to accomplish, but an easy (78.60%) one for an algorithm, $\chi^2 = 173.15$, $p < .001$. Order in which the question was asked had no impact on the distribution of the answers. Next, we looked at the perceived accuracy. The same analysis approach as in Study 2 again found a significant effect of the prediction provider, $F(1, 484) = 48.88$, $p < .001$, $dz = 0.32$. Algorithms were considered more accurate overall ($M = 4.80$; $SD = 1.39$), compared to humans ($M = 4.09$; $SD = 1.59$). There was also a main effect of response time, $F(1, 484) = 40.32$, $p < .001$, $dz = 0.29$. Slower predictions were considered more accurate overall ($M = 4.77$; $SD = 1.28$) than faster predictions ($M = 4.13$; $SD = 1.69$). More importantly, there was a significant two-way interaction, $F(1, 300) = 98.98$, $p < .001$, $dz = -0.45$. We compared the simple effects of response time on human- vs. algorithmic predictions. There was a significant effect of response time for human-generated predictions, $F(1, 242) = 165.95$, $p < .001$, $dz = 0.85$. Participants believed that slowly generated human predictions were more accurate ($M = 4.85$; $SD = 1.18$), than quickly generated predictions ($M = 3.34$; $SD = 1.59$). There was also an effect of response time for algorithm-generated predictions, $F(1, 242) = 4.74$, $p = .03$, $dz = 0.14$. In contrast to human predictions, slowly generated algorithmic predictions were seen as *less* accurate ($M = 4.69$; $SD = 1.38$) than quickly generated predictions ($M = 4.91$; $SD = 1.39$).

5.2.1. Moderated mediation model

We tested the model using STATA's GSEM builder. This was a 1–1–1 multilevel mediation model. Response time was set as the IV, effort was set as a mediator, and prediction quality evaluation was set as the DV. Crucially, prediction provider (human = −0.5 vs. algorithm = +0.5) was set as a moderator of the effort and prediction quality evaluation pathway. The overall indirect effect of perceived effort was significant, $b = 1.43$, $SE = 0.06$, $z = 25.81$, $p < .001$, 95% CI [1.32, 1.54]. However, prediction provider moderated the relationship between effort and prediction accuracy. The negative coefficient indicates a weaker relationship between effort and accuracy for algorithms, compared to humans (see upper-most section of Fig. 2).

To better understand the pattern of moderated mediation, we conducted multi-level mediations for human- and algorithmic predictions separately. For human predictions (see Fig. 2, lower left side), effort fully mediated the relationship between response time and prediction accuracy as slower response times led to the perception of more effort exerted which, in turn, led to higher prediction accuracy. For algorithms (see Fig. 2, lower right side), slower responses led to the perception of more effort exerted, but there was no relationship between effort and prediction accuracy.[6]

[5] Although this text may raise the possibility of demand effects, we note that we obtained similar results in studies that did not include this text (e.g., Study 5).

[6] We also tested the same model using perceived difficulty as the moderator instead of prediction provider. As perceived difficulty is closely related to prediction provider, we expected to obtain the same results. As expected, the results were replicated. The exact statistics are provided in the OSF materials (https://osf.io/ykamv/).

5.3. Discussion

As predicted, the asymmetric impact of different response times on the perceived accuracy of human- vs. algorithmic predictions can be explained by a mismatch in the expected difficulty of making predictions. Specifically, while making a prediction was considered to be an easy task for algorithms to accomplish, this task was seen as difficult for humans. This difference, in turn, had notable consequences in how observers responded to the inferred effort of slower response times. That is, while human effort (as inferred from slow responses) was positively correlated with the quality of another person's prediction, algorithmic effort was uncorrelated with the perceived quality of an algorithm's prediction. In the general discussion, we reflect in more detail on the implications of these findings for tasks other than predictions.

6. Study 4

In the previous study, we found that perceptions of task difficulty differed for human- vs. algorithmic predictions. In Study 4, we therefore explicitly manipulated task difficulty. Here, we expected that task difficulty would moderate the relationship between response time and perceived prediction quality. More specifically, when tasks are difficult, there should be a positive relationship between response time and quality, but when tasks are easy, there should be a negative relationship. Critically, task difficulty (rather than prediction provider) should be the primary factor that influences the relationship between response time and perceived prediction quality. In Study 4a, we use the same scenario as in Study 1 (i.e., predicting the success of students), while in Study 4b we used a different scenario. Specifically, participants had to imagine being a human resource officer predicting how long employees will be absent from work. Because the two studies had a similar procedure we again describe them together.

6.1. Method

6.1.1. Participants

Both studies were conducted on Mturk, both had 100 participants each (S4a: 39% female; $M_{Age} = 35.24$, $SD_{Age} = 11.47$; S4b: 42% female; $M_{Age} = 34.99$, $SD_{Age} = 10.00$), and the same mixed design: 2 (Prediction provider: Human vs. Algorithm; between-subject) × 2 (Response time: Fast vs. Slow; within-subject) × 2 (Task difficulty: Easy vs. Difficult; within-subject).

6.1.2. Procedure

The overall procedure was similar to Studies 1 and 2 with two differences. First, we directly manipulated the difficulty of the prediction. Participants in the easy task condition were presented with instructions which said that: "For a particular student (S4a) / employee (S4b), there were either nine or ten [one or two] valid pieces of information available, making the prediction easy [very difficult]". Second, we did not provide any numerical prediction in either of the studies.

6.2. Results

6.2.1. Perceived accuracy

A 2 (human = −0.5; algorithm = +0.5) × 2 (fast = −0.5; slow = +0.5) × (easy = −0.5; difficult = 0.5) analysis found that there was a main effect of difficulty both in S4a, $F(1, 98) = 314.78$, $p < .001$, $dz = 1.80$ and S4b, $F(1, 98) = 223.07$, $p < .001$, $dz = 1.51$. Accuracy evaluations were lower for difficult than easy predictions. There was a main effect of response time in S4b, $F(1, 98) = 5.34$, $p = .02$, $dz = 0.23$ with slowly generated predictions being judged as more accurate compared to faster predictions, but this effect did not appear in S4a.

In addition, there was a two-way interaction effect between

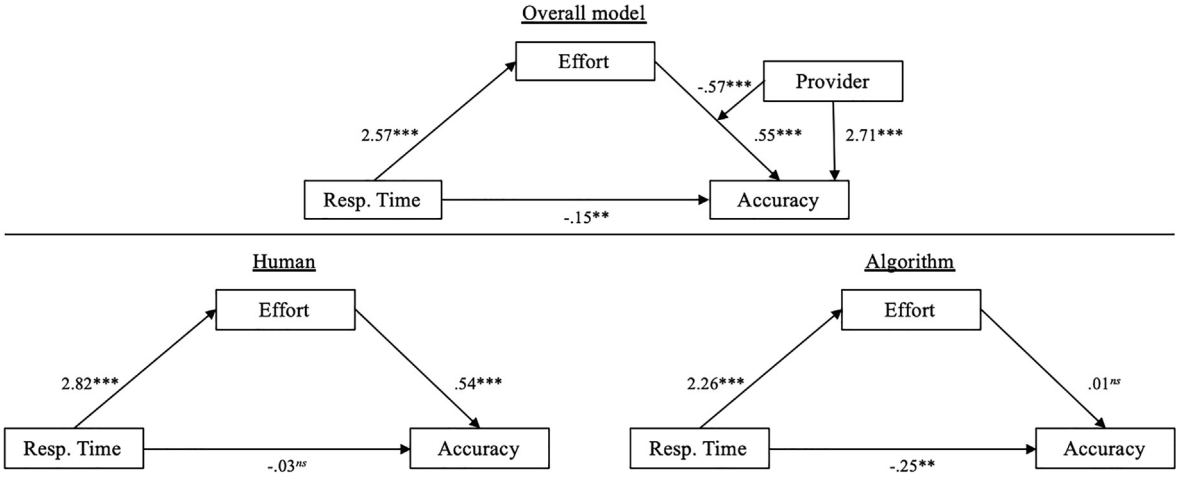

Fig. 2. Path models with corresponding coefficients for the moderated mediation model (upper section of figure), the mediation model for the human prediction provider only (lower left section of figure) and the mediation model for the algorithm prediction provider only (lower right section of the figure). *ns* $p > .05$; * $p < .05$; ** $p < .01$; *** $p < .001$. The reported coefficients are unstandardized.

response time and task difficulty both in S4a, $F(1, 98) = 20.64$, $p < .001$, $dz = 0.45$ and S4b, $F(1, 98) = 6.50$, $p = .01$, $dz = 0.26$. The interaction showed that there was a significant effect of response time for the difficult predictions both in S4a, $F(1, 99) = 6.21$, $p = .01$, $dz = 0.25$ and S4b, $F(1, 99) = 11.99$, $p = .001$, $dz = 0.35$. In S4a, there was an effect of response time for the easy predictions, $F(1, 99) = 5.58$, $p = .02$, $dz = 0.24$, but there was none in S4b. For difficult predictions, slower predictions were judged as more accurate compared to faster predictions. This reversed for the easy predictions. Slower predictions were judged as less accurate compared to faster predictions.

Finally, there was also a two-way interaction effect between prediction provider and response time both in S4a, $F(1, 98) = 12.68$, $p = .001$, $dz = 0.36$ and S4b, $F(1, 98) = 13.38$, $p < .001$, $dz = 0.37$ which showed that there was a significant effect of response time for human generated predictions both in S4a, $F(1, 48) = 21.41$, $p < .001$, $dz = 0.67$ and S4b, $F(1, 47) = 8.55$, $p = .01$, $dz = 0.43$. Just as in our previous studies, when the colleague took their time to generate the prediction, they were judged as more accurate, compared to when they were fast. However, the effect of response time was not significant for algorithmic predictions in S4a ($F = 2.13$) nor in S4b ($F < 1$), although it was in the same direction as our previous studies. Faster algorithmic predictions were judged as being of higher quality than slower ones. No other effects were significant (see Fig. 3).

6.2.2. Willingness to use

There was a main effect of difficulty both in S4a, $F(1, 98) = 168.98$, $p < .001$, $dz = 1.30$ and S4b, $F(1, 98) = 123.11$, $p < .001$, $dz = 1.11$ with more difficult predictions being less likely to be used than easier predictions. In S4a, there was also a main effect of response time, $F(1, 98) = 4.89$, $p = .03$, $dz = 0.22$ with people being less willing to use predictions that were generated fast, compared to slow. There was no effect of response time in S4b.

In addition, there was also a two-way interaction effects between response time and difficulty both in S4a, $F(1, 98) = 13.75$, $p < .001$, $dz = 0.37$ and S4b, $F(1, 98) = 5.71$, $p = .02$, $dz = 0.24$ which showed that there was a significant effect of response time for the difficult predictions both in S4a, $F(1, 99) = 13.82$, $p < .001$, $dz = 0.37$ and S4b, $F(1, 99) = 4.05$, $p = .05$, $dz = 0.20$, but there was no effect for easy predictions in either study. For difficult predictions, people were more willing to use slower compared to faster generated predictions.

Finally, there was also a two-way interaction between prediction provider and response time both in S4a, $F(1, 98) = 13.02$, $p < .001$, $dz = 0.36$ and S4b, $F(1, 98) = 15.19$, $p < .001$, $dz = 0.39$ which showed that there was a significant effect of response time on human-generated predictions in S4a, $F(1, 48) = 6.28$, $p = .02$, $dz = 0.39$, and in S4b, $F(1, 47) = 5.81$, $p = .02$, $dz = 0.35$. Again, when the colleague took their time to generate a prediction, participants were more likely to use it than when they were fast. However, there was no significant effect of response time on algorithmic predictions in S4a ($F < 1$) nor in S4b ($F = 1.74$) although they were in the same direction as previous studies, with participants saying that they were more likely to use them for fast predictions than slow predictions. No other effects were significant.

6.3. Discussion

The results of both Study 4a and 4b show that once difficulty is explicitly manipulated, response time has a similar effect on the perceived accuracy of predictions for both algorithms and humans. Critically, task difficulty moderated the relationship between different response times and prediction quality: when the task was difficult, there was a positive relationship between response time and quality, but when the task was easy there was a negative relationship.

7. Study 5

In the previous studies, we relied on a within-subjects manipulation of response time. We focused on this approach because decision-makers often have repeated encounters with the same person or algorithmic support system. Nevertheless, it could be that response time is a much more easily evaluable attribute for humans as compared to algorithms (Hsee & Zhang, 2010). Arguably, the average person has more prior experience with human predictions than algorithmic predictions, and this lack of experience with algorithms may make it more difficult to evaluate changes in an algorithm's response time. In Study 5, we therefore focus on algorithms and test the effect of response time on prediction quality evaluations in a single (between-subject) evaluation design. Crucially, we expected the effect of response time to become stronger once participants experienced *multiple* fast or slow algorithmic predictions.

7.1. Method

7.1.1. Participants

Two-hundred and forty-one participants were recruited on Prolific. The study had a single between-subject factor of response time (Fast vs. Slow). After excluding the people who failed an attention check presented at the end of the study, we were left with 236 participants (60%

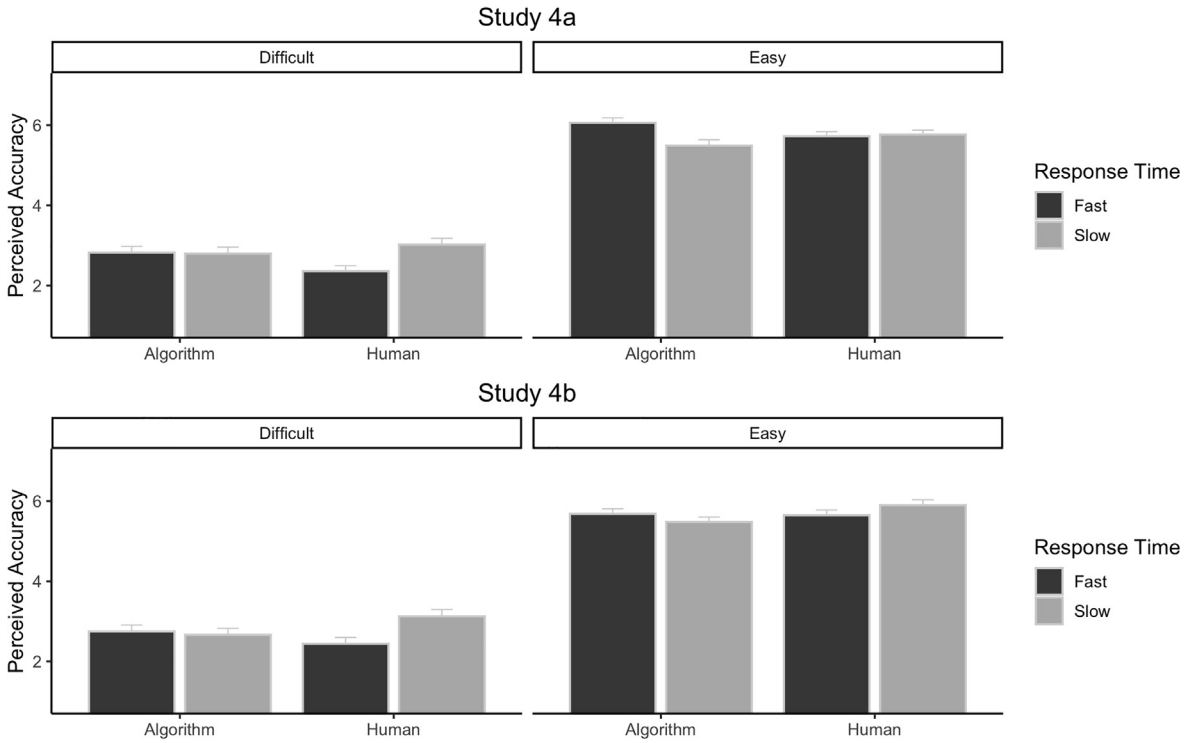

Fig. 3. The means and standard errors of Study 4a (upper row) and Study 4b (lower row) results on perceived accuracy of the generated prediction as a function of prediction provider (Algorithm vs. Human), response time (Fast vs. Slow), and task difficulty (Easy vs. Difficult).

female; M_{Age} = 35.44, SD_{Age} = 11.91).

7.1.2. Procedure

We used a realistic task where participants were presented with English Championship League football predictions for an upcoming round of matches. We chose the Championship League, rather than the Premier League (which has some of the most famous teams in the world, e.g., Manchester United, Liverpool, etc.) to avoid our participants being too familiar with the task – in which case they may disregard algorithmic predictions entirely. The predictions presented to the participants were made by an actual algorithm from the "FiveThirtyEight" website.

Participants evaluated the quality of 12 predictions made by an algorithm called "StatCast". The league has 24 teams; hence 12 matches and 12 predictions were made for each weekly round of matches. Participants were told that the algorithm was developed at the Eindhoven University of Technology to predict the outcome of sporting matches. The presented matches were scheduled one week after we collected the data for this study. To expand on our main dependent variable, for each match, participants were asked: "How accurate do you think is StatCast's prediction?", and "How persuasive do you think is StatCast's prediction?" Ranging from −3 (*Not at all*) to 3 (*Very much*). To describe the predictions, we used the same wordings from previous studies. For fast predictions, we added: "Instantly", "Quite rapidly", and "With little or no delay". For slow predictions, we added: "With a substantial lag", "After a lengthy period", and "After an extensive delay". We had six response time wordings for both fast- and slow speeds so the wordings were shown twice each, given that we had 12 trials. At the end, after going through all 12 trials, participants we asked if they were a fan of any particular club within the league (if they said yes, they were asked to type in the name of the club).

7.2. Results

The two measures of accuracy and persuasiveness were highly correlated, $r = 0.76$, $p < .001$ so we made one composite measure of perceived prediction quality (by averaging the answers). We first verified whether, taking into account all 12 trials, we would observe the same effect of response time as in our previous studies. Note that now, participants were presented with the *same* response time descriptions, i.e., either just fast, or just slow. As expected, there was an effect of response time, $F(1, 234) = 15.58$, $p < .001$, $dz = 0.26$. Prediction quality in the slow condition was judged as lower ($M = 4.06$, $SD = 1.44$) than in the fast condition ($M = 4.68$, $SD = 1.48$).[7]

Subsequently, we tested the effect of response time solely for first trials. We observed the same effect of response time, $F(1, 234) = 6.03$, $p = .01$, $dz = 0.16$ although considerably smaller than the overall effect ($M_{slow} = 4.33$; $M_{fast} = 4.77$). As expected, when we looked at the effect of response time solely for the last trials that participants experienced, the same effect was present, although much larger, $F(1, 234) = 16.35$, $p < .001$, $dz = 0.26$ ($M_{fast} = 4.74$; $M_{slow} = 3.96$). More experience with the same algorithm thus increased participants' sensitivity to algorithmic response times. Looking across all 12 trials, we see that predictions with slower responses were evaluated as worse over time (see Fig. 4).

7.3. Discussion

Relying on sports predictions, we successfully replicated the same effect of algorithmic response times, but now in a between-subjects design. Specifically, participants who only experienced slowly generated predictions by an algorithm judged these predictions as worse than those who only experienced fast predictions. The effect increased as participants' experience with the algorithm increased.[8] Slow

[7] Twelve participants said that they were a fan of a specific club in the league. Excluding those participants, the effect remained significant and was slightly stronger at $dz = 0.27$.

[8] We also looked at how people evaluate prediction advice quality independent of seeing all other response time manipulations in *all the other studies*

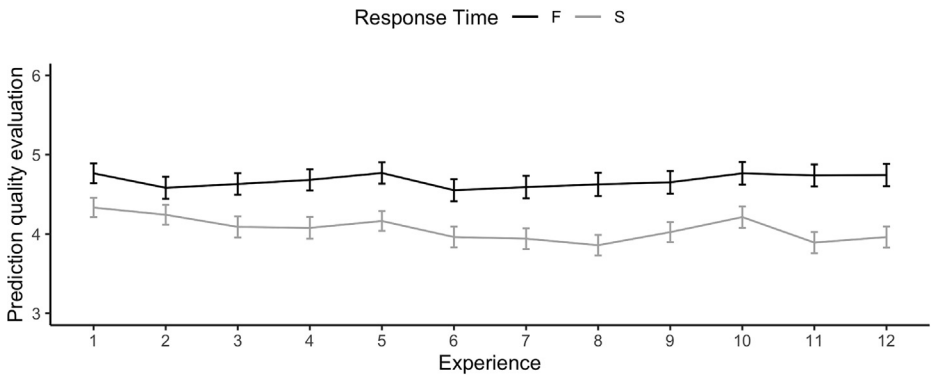

Fig. 4. The means and standard errors of Study 5 on advice quality as a function of response time (Fast vs. Slow) and experience with the algorithm (i.e., ranging from the first to the twelfth trial).

predictions were evaluated as much worse on the later trials, while the quality evaluations for fast predictions remained relatively stable over time. Experience with an algorithm is thus an important moderator of the effect of different algorithmic response times on people's quality evaluations.

8. Study 6

In the last two studies, we extend our findings to *behavioral consequences* of observing slow- vs. fast algorithmic predictions. We focused solely on algorithms, as people are particularly unwilling to use algorithm-generated advice, which is often better than advice generated by humans (Carroll et al., 1982; Dietvorst et al., 2015; Önkal et al., 2009). This means that not following algorithm-generated advice can have potentially negative consequences. In Study 6, we looked at the consequences of different algorithmic response times on seeking additional advice beyond the one provided by an algorithm. We expected that participants presented with slow (vs. fast) algorithmic predictions would be more likely to choose to use a human-generated prediction instead. In addition, we recruited a separate (smaller) sample of participants to gauge how willing people would be to go to another human prediction provider, where no information about the algorithm's response time was provided. We hoped that this would help us to position the effect more clearly (i.e., identify if the effects of different response times were driven more by slow- or fast algorithmic predictions).

8.1. Methods

8.1.1. Participants

Two hundred and twenty-six participants were recruited on MTurk. There was a single within-subject condition of response time. After excluding participants who did not pass the initial attention check and those who did not complete the full study, we had 200 participants (42% female; M_{Age} = 35.89, SD_{Age} = 11.28). Simultaneously, an additional 63 participants were recruited for the separate "no response time info" condition. After excluding those who did not pass the attention check and those who did not complete the full study, we were left with 50 participants in this condition (50% female; M_{Age} = 34.24, SD_{Age} = 9.16).

8.1.2. Procedure

The procedure was similar to Study 2. The only difference was the wording of the main dependent variable which now read: "Given StatCast's response time, how likely are you to disregard its prediction and consult a colleague instead" – ranging from −3 (*very unlikely*) to 3 (*very likely*).

8.2. Results[9]

Willingness to disregard the algorithmic prediction. As expected, our analyses indicated that people were more likely to disregard the algorithm's prediction for a colleague's when it was generated slowly (M = 3.90, SD = 1.66) as opposed to quickly (M = 3.48, SD = 1.96), F (1, 199) = 7.15, p = .01, dz = 0.20.

No info about response time. When no information about the algorithm's response time was given, the average willingness to consult a colleague was similar to the fast condition (No information: M = 3.54, SD = 1.83; Fast prediction: M = 3.48; SD = 1.96; t(248) = 0.20, p = .84). These results indicate that the effect of response time is most likely driven by situations when the algorithm took its time to generate the prediction.

8.3. Discussion

Results of Study 6 show that the effect of different algorithmic response times extends to situations where participants are given an opportunity to consult another person for a prediction. People were more likely to disregard slow (vs. fast) algorithmic predictions.

9. Study 7

In our final study, we conducted an incentivized test of the behavioral consequences of observing algorithmic response times, relying on the sports prediction task introduced in Study 5. Specifically, participants were given the opportunity to choose those sports predictions that would go towards a monetary bonus. That is, we paid an extra reward

(*footnote continued*)
we use the within-subject manipulation of response time. We focused only on the first trial that participants saw (i.e., either a single fast or a single slow prediction). We found that, for humans, the same effect of response time can be observed. i.e., slower predictions were judged as being of higher quality. For algorithms, however, there was no difference, i.e., simply seeing either one fast or one slow prediction generated by an algorithm, did not have an effect on prediction advice quality. This is consistent with our proposition that response time is a more evaluable attribute for humans, than algorithms. For more detail about the analysis please see the supplementary material.

[9] Participants were also asked to evaluate how much effort they thought the algorithm exerted. Slower predictions were again evaluated as the algorithm exerting more effort, F(1, 199) = 99.56, p < .001, dz = 0.71. In addition, at the end of the study, participants were also asked to evaluate StatCast's quality as an algorithm given the time it took to provide the predictions, evaluating all six different response time descriptions. The graphical representation of the answers essentially indicates that StatCast was judged as being of lower quality for slow speed descriptions. The analysis code allows the interested reader to generate the graph, but we do not consider it relevant to report it in the main text of the article.

for each prediction that the participant chose and that turned out to be true (e.g., if the algorithm suggested Blackburn Rovers would win and they actually won, participants would get an extra £0.05). Data were collected two days before the first match was scheduled. We hypothesized that people would be more likely to choose a sports prediction that the algorithm generated fast as opposed to slow. In addition, we also wanted to explore whether there would be any differences between a UK sample (which should be more familiar with the English Championship League) and a US sample (which should be less familiar with it) in how different response times would impact quality evaluations and behaviors.

9.1. Method

9.1.1. Participants

Three hundred and forty-nine people took the survey on Prolific. After excluding people who failed the attention check or simply did not complete the full study, we were left with 200[10] participants (60% female; M_{Age} = 34.66, SD_{Age} = 11.81). The sample had 100 participants from the UK (72% female; M_{Age} = 35.48, SD_{Age} = 11.68) and 100 participants from the US (48% female; M_{Age} = 33.84, SD_{Age} = 11.95). Response time (Fast vs. Slow) was the only within subject factor.

9.1.2. Procedure

The procedure was similar to Study 5 but for five differences. First, the matches were updated to select upcoming matches at the time that this study was conducted. Second, response time was provided in actual numbers to participants. Specifically, for each trial, a random number ranging from 4.9 to 6.9 was generated. In the fast conditions, 4 s were subtracted from this number while in the slow conditions, 6 s were added to illustrate the algorithm's response time. This way, we also knew which response time each participant saw. Third, after going through the 12 trials, participants were shown a list of all the predictions with the same response times that they saw during the trials. They could then choose three of these predictions as "their own", meaning that they would receive an additional monetary reward of £0.05 for each of the predictions that turned out to be true. There was no deception involved since we verified the results after the matches were played and paid out each participant dependent on their choices. Fourth, towards the end, we explored participants' familiarity with the English Championship League by presenting them with four statements for which they had to indicate their agreement from −3 (*Completely disagree*) to 3 (*Completely agree*). The statements were: "I am an avid fan of the English Championship League", "I consider myself an expert when it comes to the English Championship League", "I watch at least one of the English Championship League matches every week (during the season)", "I am familiar with the current standings in the English Championship League." Cronbach's alpha was very high at 0.94 so we made one composite measure by averaging the results of the four statements. Fifth, for each prediction (i.e., each match), it was randomly determined whether StatCast predicted the outcome of the match in a fast or slow way.

9.2. Results

The two measures of accuracy and persuasiveness were highly correlated, r = 0.80, p < .001 so we made one composite measure of perceived prediction quality (by averaging the answers). Consistent with previous studies, we found a significant effect of response time, F(1, 199) = 29.95, p < .001, dz = 0.39. Participants considered slow algorithmic predictions to be of a lower quality (M = 4.05; SD = 1.54), compared to fast predictions (M = 4.84; SD = 1.63).

To verify whether there were any differences in familiarity between UK and US participants, we compared our participants' scores on the familiarity measure. Indeed, participants in the UK said that they were more familiar with the English Championship League (M = 2.61; SD = 1.71) than participants in the US (M = 1.58; SD = 1.10), t(198) = 5.05, p < .001, dz = 0.72. Including country as a variable in our analysis, we again obtained an effect of response time, F(1, 198) = 30.54, p < .001, dz = 0.40, and a two-way interaction with country and response time, F(1, 198) = 5.30, p = .02, dz = 0.16. There was no main effect of country (F < 1). In decomposing the interaction, we found a significant effect of response time for both the UK, F(1, 99) = 6.53, p = .01, dz = 0.26 and US participants, F(1, 99) = 24.76, p < .001, dz = 0.50, although it is clear that the difference in quality evaluations for predictions made quickly and predictions made slowly was much stronger for US participant as compared to UK participants (see Fig. 5).

We also verified whether there would be an effect of response time if we did not use the categorical (Fast vs. Slow) conceptualization as the independent variable, but instead if we used the actual numerical values of response times shown to the participants. Again, there was a clear negative relationship b = −0.14, SE = 0.054, t(1607.6) = −2.50, p = .01, indicating that the longer it took an algorithm to come to a prediction, the lower the perceived quality of its prediction.

Choice data. Each person could choose three predictions that would go towards their bonus, meaning 600 choices were made in total. Had people shown no preference for either fast or slow predictions, we would have observed something close to a 50/50 distribution. However, and in accordance with our expectations, the data showed that people actually chose 381, or 63.5% fast predictions overall. A binomial test indicated that this was significantly different than the expected 50/50 distribution, p < .001 (two-sided). Looking only at UK participants, 59.3% of their choices favored a fast prediction. A binomial test again indicated that this was significantly different from the 50/50 distribution, p = .001 (two-sided). As expected, for US participants, even more choices favored fast predictions (67.6%), p < .001 (two-sided).

9.3. Discussion

Using sports predictions, a more concrete response time manipulation (i.e., using numbers rather than textual descriptions), and an incentivized prediction task, we confirmed that slow response times had a detrimental impact on the perceived quality of algorithmic prediction. People judged slower predictions as less accurate and less persuasive, and they were less likely to rely on them for their bonuses. This tendency was much more pronounced in the US sample, where familiarity with the English Champions League (the domain in which the predictions were made) was much lower. Thus, response time was a much more relied upon cue in situations that are unfamiliar, leading individuals to display an even stronger condemnation for slowly generated algorithmic predictions.

10. General discussion

When are people reluctant to trust algorithmic predictions? Here, we demonstrate that it depends on the algorithm's response time. People judged slowly (vs. quickly) generated predictions by algorithms as being of lower quality. Further, people were less willing to use slowly generated algorithmic predictions. For human predictions, we found the opposite: people judged slow human-generated predictions as being

[10] In our preregistration, we stated that we would exclude participants that spent, on average, more than 10 s on each trial as this might indicate that they have looked up information about the games. After verifying the average times, we realized we underestimated the necessary time as 98 participants would need to be excluded. We decided to void this aspect of our registration since it would mean discarding 50% of our sample resulting in a serious lack of statistical power to detect an effect.

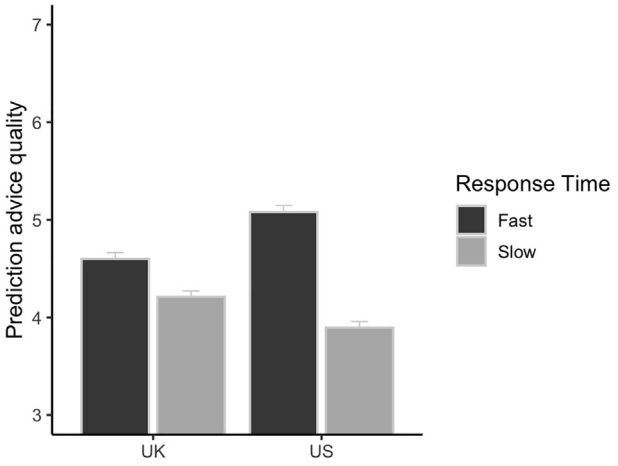

Fig. 5. The means and standard errors of Study 7 results on advice quality as a function of participants' country of origin (UK vs. US) and response time (Fast vs. Slow).

of higher quality. Similarly, they were more likely to use slowly generated human predictions.

We find that the asymmetric effects of response time can be explained by different expectations of task difficulty for humans vs. algorithms. For humans, slower responses were congruent with expectations; the prediction task was presumably difficult so slower responses, and more effort, led people to conclude that the predictions were high quality. For algorithms, slower responses were incongruent with expectations; the prediction task was presumably easy so slower speeds, and more effort, were unrelated to prediction quality. In short, response times have a nuanced effect on advice quality evaluations. Indeed, for more difficult judgments, longer response times may lead to similar perception of quality for algorithms as for humans, namely: slower responses leading to higher quality evaluations.

Similarly, we find that the effect of algorithmic response times on prediction quality evaluations appeared both in a between- and within-subject setting, and that the effect of response time is moderated by a person's experience with an algorithm. Specifically, as people repeatedly experienced slow algorithms, the (detrimental) effect of slow algorithmic responses on prediction quality evaluations became stronger. Finally, focusing on algorithms specifically, we find that slow algorithmic predictions can lead people to seek out additional advice from other humans. Confirming the importance of response time as a cue, a subset of people who were unfamiliar with the prediction domain relied even more on the time algorithms needed to make predictions.

Previous research has identified response time as an important cue in social interactions (Critcher et al., 2013; Evans & van de Calseyde, 2017; Mata & Almeida, 2014; Van de Calseyde et al., 2014) and participants in our studies also used it as information to evaluate the quality of others' predictions. However, while most prior research indicates that observed response times are interpreted in terms of doubt (Critcher et al., 2013; Evans & van de Calseyde, 2017; Van de Calseyde et al., 2014), the current results demonstrate that response times can also be interpreted in terms of effort (Jago & Laurin, 2018; Kupor et al., 2014). More specifically, if doubt (rather than effort) was the main information that response times signaled, we would have seen different results. That is, people would have perceived fast predictions by others as more accurate as faster response times have been shown to indicate more confidence (Van de Calseyde et al., 2014) and people generally prefer confident (over doubtful) predictions (Stavrova & Evans, 2019).

Interestingly, while people interpreted algorithmic response times in terms of effort (i.e., slow predictions indicate more effort exertion by an algorithm), people seem to see it as undiagnostic when evaluating the quality of predictions. We speculate that this is due to the fact that algorithms are judged more as tools that perform complicated tasks following closed and structured procedures (Simon & Neisser, 1992). Therefore, tasks that involve complex calculations are seen as easy for algorithms to accomplish, making the presence or absence of effort relatively meaningless. Nonetheless, while perceived effort did not serve as a suitable mechanism in explaining how algorithmic response times affect quality evaluations, there could be other possible mechanisms that govern this relationship. One potential avenue for future research is to investigate whether people have default assumptions about algorithms such that observing slowness might be indicative of an algorithm's "bugginess".

The model that relies on task difficulty as a moderator of response times allows for several predictions that are relevant for future research. For instance, following this model, we would predict that tasks that are seen as difficult (easy) for algorithms (humans), slower response times would lead to higher (lower) quality evaluations. This theorizing is also relevant to other domains such as moral judgments. Previous work suggests that increased deliberation on tragic trade-offs reaffirms the solemnity of the occasion (i.e., longer response times breed trust), while deliberation on taboo trade-offs undermines trust (Tetlock, Kristel, Elson, Green, & Lerner, 2000). Thus, in some cases, the longer one takes on contemplating indecent proposals, the more one's moral identity is compromised. It could be that moral judgments constitute a separate cognitive arithmetic and are thus differently amenable to response times than other judgments (e.g., forecasting, recognition, calculation). It is worth pointing out that recent evidence suggests that people seem to be strongly averse to algorithms making any sort of moral decisions (Bigman & Gray, 2018), so a challenge for future research is to understand how response time might modulate trust in algorithmic advice when applied to the moral domain.

Response time also seems to be a more evaluable attribute for humans than for algorithms. We obtained several indications for this notion throughout our studies. First, effect sizes of response time for humans were consistently much larger than for algorithms. Second, the response time effect was reliably obtained for humans even when experiencing only a single indication of fast or slow response time (i.e., a between-subject design – see also supplementary material). Conversely, for algorithms, it appears that experience with the algorithm can play a crucial role as the results of Study 5 suggest. It is worth pointing out thought that Study 5 did not include a human prediction provider condition which would have allowed for a direct comparison of between-subject effects across both human and algorithm predictions providers. Consistent with general evaluability theory (Hsee & Zhang, 2010), people might not have relevant reference information for different response times in algorithms. As it increasingly becomes more likely that people will interact with the same algorithms, sensitivity to the attribute of response time might play an important role in how we evaluate algorithm-generated advice in the future.

In our studies, people were generally trusting of algorithms – predictions provided by algorithms were judged to be better overall. These results are in line with the idea that algorithm aversion primarily arises when people observe an algorithm fail (Dietvorst et al., 2015; Dietvorst, Simmons, & Massey, 2016). Similarly, other recent work has found that advice has a greater impact on people when they think it comes from algorithms (Logg, Minson, & Moore, 2019) and the reported findings in the current paper are consistent with this notion.

Practically, our results could have important implications: algorithmic response times can have a profound impact on the way people evaluate and use advice. This implies that people might be sensitive to imperfections, glitches, or delays, when advice by an algorithm is being provided, leading them to adversely (and perhaps erroneously) disregard the advice – in particular when people have repeated experiences with an algorithm. As already argued, this could have various negative consequences such as leading people to solicit further advice or, if the advice situation is particularly unfamiliar, a larger reliance on response time as a cue. Conversely, making fast response times salient may increase a person's reliance on algorithmic predictions. Future

research could address this interesting question in more detail by testing whether and when response times can be used as a nudge to increase a person's trust in algorithmic advice.

In the supplementary material, we report an additional two studies that tackle the question whether prediction provider's expertise, and the direction of the prediction (i.e., whether an increase or a decrease was predicted) moderate the impact of different response times on human- vs. algorithmic predictions. Study 8 looked at the potential impact of advice provider expertise. For average expertise, both human- and algorithmic predictions were considered more accurate when provided slowly, compared to predictions provided quickly. However, we observed no effects in the expert conditions, possibly due to a ceiling effect. Finally, Study 9 focused only on algorithmic predictions and looked at whether response time would have a different impact dependent on whether the prediction was of an increase compared to a decrease. Prediction direction did not have an effect. Another important direction for future research is to look at situations which are inherently riskier, more important in terms of their consequences, and more high-stakes. While general algorithm aversion could apply for these situations (Logg, 2017) and it seems that people still have misgivings on applying algorithms in such situations, important cues like response time could moderate algorithm advice evaluation.

11. Conclusion

Given the ubiquity of prediction algorithms, as well as their general superiority in providing high-quality advice, understanding how subtle cues may impact the way people evaluate algorithms is both timely and important. The present research is an initial step towards understanding this matter by demonstrating how different algorithmic response times affect people's evaluations and behaviors. A very simple cue such as response time, which at times can even be just a random fluctuation, can evidently lead individuals to disregard or adopt an algorithm's solution.

CRediT authorship contribution statement

Emir Efendić: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Philippe P.F.M. Van de Calseyde:** Conceptualization, Methodology, Funding acquisition, Investigation, Writing - original draft, Writing - review & editing. **Anthony M. Evans:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

Appendix A

OSF link to data, materials, and analysis code: https://osf.io/ygeha/
Links for preregistrations of individual studies:

Study 1: https://osf.io/9esdv/
Study 2: https://osf.io/yrhjn/
Study 3: https://osf.io/m48wq
Study 4b: https://osf.io/tj562/
Study 6: https://osf.io/ebk6h/
Study 7: https://osf.io/s8rbd/

Supplementary material data and analysis code: https://osf.io/qsdbz/.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2020.01.008.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1)https://doi.org/10.18637/jss.v067.i01.

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., ... Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine, 3*(108), https://doi.org/10.1126/scitranslmed.3002564 108ra113-108ra113.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003.

Carroll, J. S., Wiener, R. L., Coates, D., & Galegher, J. (1982). Evaulation, Diagnosis, and prediction in parole decision making. *Law & Society Review, 17*, 199.

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.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*(3), 308–315. https://doi.org/10.1177/1948550612457688.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26*(2), 180–188. https://doi.org/10.1037/h0030868.

Diab, D. L., Pui, S.-Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment, 19*(2), 209–216. https://doi.org/10.1111/j.1468-2389.2011.00548.x.

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.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: people will use imperfect algorithms if they can (Even Slightly) modify them. *Management Science, mnsc.2016.2643*. https://doi.org/10.1287/mnsc.2016.2643.

Evans, A. M., & van de Calseyde, P. P. F. M. (2017). The effects of observed decision time on expectations of extremity and cooperation. *Journal of Experimental Social Psychology, 68*, 50–59. https://doi.org/10.1016/j.jesp.2016.05.009.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford University Press.

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. https://doi.org/10.1287/inte.1070.0309.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science, 5*(4), 343–355. https://doi.org/10.1177/1745691610374586.

Jago, A. S., & Laurin, K. (2018). Inferring commitment from rates of organizational transition. *Management Science*. https://doi.org/10.1287/mnsc.2017.2980.

Konovalov, A., & Krajbich, I. (2017). Revealed Indifference: Using Response Times to Infer Preferences (SSRN Scholarly Paper No. ID 3024233). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3024233.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25 (pp. 1097–1105). Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology, 40*(1), 91–98. https://doi.org/10.1016/S0022-1031(03)00065-9.

Kupor, D. M., Tormala, Z. L., Norton, M. I., & Rucker, D. D. (2014). Thought calibration: How thinking just the right amount increases one's influence and appeal. *Social Psychological and Personality Science, 5*(3), 263–270. https://doi.org/10.1177/1948550613499940.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13)https://doi.org/10.18637/jss.v082.i13.

Labroo, A. A., & Kim, S. (2009). The "instrumentality" heuristic: Why metacognitive difficulty is desirable during goal pursuit. *Psychological Science, 20*(1), 127–134. https://doi.org/10.1111/j.1467-9280.2008.02264.x.

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.

Logg, J. M. (2017). Theory of Machine: When Do People Rely on Algorithms? Harvard Business School Working Paper Series # 17-086. Retrieved from https://dash.harvard.edu/handle/1/31677474.

Mata, A., & Almeida, T. (2014). Using metacognitive cues to infer others' thinking. *Judgment & Decision Making, 9*(4), 349–359.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. https://doi.org/10.1037/11281-000.

Norton, M. I., Mochon, D., & Ariely, D. (2012). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology, 22*(3), 453–460. https://doi.org/10.1016/j.jcps.2011.08.002.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown/Archetype.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*(4), 390–409. https://doi.org/10.1002/bdm.637.

Porter, J. (2018). Robot lawyer DoNotPay now lets you 'sue anyone' via an app. Retrieved November 15, 2018, from The Verge website: https://www.theverge.com/2018/10/10/17959874/donotpay-do-not-pay-robot-lawyer-ios-app-joshua-browder.

Power, D. J. (2002). *Decision support systems: Concepts and resources for managers.* Greenwood Publishing Group.

Simon, H., & Neisser, U. (1992). Can computers help us understand the human mind. In Taking sides: Clashing views on contraversial psychological issues (pp. 128–143).

Stacey, D., Légaré, F., Lewis, K., Barry, M. J., Bennett, C. L., Eden, K. B., ... Trevena, L. (2017). Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews, 4*. https://doi.org/10.1002/14651858.CD001431.pub5.

Stavrova, O., & Evans, A. M. (2019). Examining the trade-off between confidence and optimism in future forecasts. *Journal of Behavioral Decision Making, 32*, 3–14.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*(5), 853–870. https://doi.org/10.1037/0022-3514.78.5.853.

Van de Calseyde, P. P. F. M., Keren, G., & Zeelenberg, M. (2014). Decision time as information in judgment and choice. *Organizational Behavior and Human Decision Processes, 125*(2), 113–122. https://doi.org/10.1016/j.obhdp.2014.07.001.

Wachter-Boettcher, S. (2017). *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech.* W. W. Norton & Company.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014.

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society, 20*(1), 137–150. https://doi.org/10.1080/1369118X.2016.1200645.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112.