

Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making

YOYO TSUNG-YU HOU, Information Science, Cornell University, USA

MALTE F. JUNG, Information Science, Cornell University, USA

The increased use of algorithms to support decision making raises questions about whether people prefer algorithmic or human input when making decisions. Two streams of research on algorithm aversion and algorithm appreciation have yielded contradicting results. Our work attempts to reconcile these contradictory findings by focusing on the framings of humans and algorithms as a mechanism. In three decision making experiments, we created an algorithm appreciation result (Experiment 1) as well as an algorithm aversion result (Experiment 2) by manipulating only the description of the human agent and the algorithmic agent, and we demonstrated how different choices of framings can lead to inconsistent outcomes in previous studies (Experiment 3). We also showed that these results were mediated by the agent's perceived competence, i.e., expert power. The results provide insights into the divergence of the algorithm aversion and algorithm appreciation literature. We hope to shift the attention from these two contradicting phenomena to how we can better design the framing of algorithms. We also call the attention of the community to the theory of power sources, as it is a systemic framework that can open up new possibilities for designing algorithmic decision support systems.

CCS Concepts: • Human-centered computing → Collaborative and social computing → Empirical studies in collaborative and social computing

KEYWORDS: decision support systems; decision aids; augmented decision making; algorithm aversion; algorithm appreciation; expert power; human-algorithm interaction

ACM Reference format:

Yoyo T.-Y. Hou and Malte F. Jung. 2021. Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5, CSCW2, Article 477 (October 2021), 25 pages, <https://doi.org/10.1145/3479864>

1 INTRODUCTION

With the advance of artificial intelligence (AI) technology, algorithmic decision support systems (or decision aid, augmented decision making, expert systems [5]) increasingly facilitate people's decision making processes by providing information, suggestions, or candidates. Recommendation systems help with decision making by providing consumers with algorithm-selected items, lowering cognitive load in navigating through millions of options. Resume screening systems automatically summarize each candidate's file with a score so that recruiters

Authors' email addresses: th588@cornell.edu, mjf28@cornell.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

2573-0142/2021/10 – ART477 \$15.00

© Copyright is held by the owner/author(s). Publication rights licensed to ACM.

<https://doi.org/10.1145/3479864>

do not have to read through all applications to identify the most desirable candidates. In hospitals, expert systems help health professionals by suggesting possible interpretations in patient's examination reports and imaging. In many situations, these systems take decision-support roles traditionally held by co-workers or advisors, competing directly with human intelligence. These situations pose a pressing question: when making decisions, are people more influenced by input from algorithms or from humans?

This seemingly straight-forward question, however, has generated two streams of studies with contradicting results. Early studies of algorithm-support in decision making suggested that people tend to dismiss input from algorithms even when given information about the algorithm's superior performance—a phenomenon called “algorithm aversion” [8]. Algorithm aversion has been found in studies across different scenarios and knowledge domains [3,7,32], and the concept's popularity even led to its adoption in public media [12]. Despite the growing evidence for algorithm aversion, findings from recent studies suggested that an opposite response to algorithms is possible: in some situations, people rely more on algorithmic advice than human advice, a phenomenon called “algorithm appreciation” [20]. These two streams of research, algorithm aversion and algorithm appreciation, predict different outcomes in situations where people receive decision inputs from algorithms and humans. They also provide different suggestions as to how to increase the acceptability of algorithmic decisions, forming contradictions that remain unresolved. While many studies have investigated the reasons and factors influencing algorithm aversion, such as task objectivity [7] and people's prior perception of algorithms [8], few have addressed the relationship between algorithm aversion and algorithm appreciation. It is unclear whether these two phenomena are independent or whether there is a common factor determining between them.

The purpose of this paper is to reconcile the seeming conflict between literature on algorithm aversion and algorithm appreciation. Through three studies, we demonstrated framing, i.e., the way in which an algorithmic agent and a human agent were introduced, as a potential mechanism. We found that the framing of algorithmic and human decision aids will influence their perceived competence, i.e., expert power, which will in turn influence whether people adhere more to the algorithm's input (algorithm appreciation) or the human's input (algorithm aversion). We showed this by two almost identical decision making experiments: by manipulating only the descriptions of the algorithm and the human, we can create an algorithm appreciation result (Experiment 1) as well as an algorithm aversion one (Experiment 2). We then further developed this idea by comparing different framings of human agents and algorithmic agents, showing how different choices of framings might lead to inconsistent findings in previous studies (Experiment 3). We also demonstrated through mediation analyses that expert power was an important factor across all three experiments, suggesting that we can make agents more influential by framing them with more expert power.

Our paper makes three contributions to the literature on AI-supported decision making: First, by identifying that framing is a key mechanism behind the differentiation of algorithm aversion and algorithm appreciation, our work advances research on this topic toward a theory that integrates literature on these two phenomena. We also hope, by this integration, to direct the debate between algorithm aversion and algorithm appreciation to how we can better design the framing of algorithms. Second, although there is increasing interest in CSCW on AI-supported decision making [6,14,15,19], the CSCW literature has paid little attention to growing work on algorithm aversion and algorithm appreciation. We therefore contribute to the CSCW community by bring in this pressing question as well as a refined version of a useful research paradigm (the

judge–advisor system paradigm) on this topic. Third, we call for the attention of the community to the theory of power sources in Organizational Behavior, as we believe that this framework brings in a systemic approach for increasing the influence of algorithmic decision support systems and, more generally, for designing cooperative human-agent interaction.

2 LITERATURE REVIEW

2.1 Algorithm Aversion & Algorithm Appreciation

In the effort of knowing how people use decision support systems, early studies have shown that human decision makers are reluctant to trust the selections or suggestions from algorithms compared to those from humans, even when algorithms exhibit superior performance. This phenomenon is called “algorithm aversion.” It was first coined in studies that compared the influence of algorithm’s suggestions against the experiment participant’s own opinions [8,9], but subsequent studies have generated a stream of similar results when comparing algorithms’ versus other people’s suggestions [3,7,32], showing that algorithm aversion does not simply result from people’s overconfidence in their own reasoning. It has also been found across domains such as investment [23], medicine [21,25], and content recommendations [13]. The fact that people disregard suggestions from algorithms highlights a barrier in the adoption of algorithmic decision support systems, which have led many researchers to investigate its causes and to find ways to alleviate it. For example, studies have shown that algorithm aversion becomes more pronounced after seeing an algorithm make a mistake [8]. However, this effect can be alleviated if users can modify how the algorithm works [9] or if they believe that algorithm can learn [2]. Studies also suggest that algorithm aversion is most severe when the tasks are perceived as subjective instead of objective, or when the algorithm is perceived to have low human-likeness [7]. These findings have drawn much attention across different academic fields and have yielded two literature review papers [5,18] and even public media coverage on algorithm aversion [12].

On the other hand, a growing number of studies have observed an opposite phenomenon, that humans are influenced more by decision inputs from algorithms compared to those from humans. The term “algorithm appreciation” was first proposed by Logg et al. [20], who showed in their study that, unlike the “received wisdom”, people are actually influenced more by advice when they think it is from an algorithm than from humans, regardless of the subjectivity and objectivity of the tasks. This finding of algorithm appreciation echoes findings in previous studies on image analysis [31] and online saving systems [16] which shows that participants adhered more to algorithmic input than human input. It has also stimulated many subsequent studies that start investigating algorithm appreciation [4,28], with findings suggesting a similar preference for algorithmic decision making in other domains such as public health [1].

In sum, research on algorithm aversion and algorithm appreciation gives contradicting answers to the question whether people prefer an algorithm’s or a human’s decision input, and the contradicting findings have left three challenges to be resolved. First, there is confusion about the factors that influence the acceptance of algorithms. For example, in terms of the task type, Castelo et al found that when the task is perceived as more objective, the algorithm aversion is less severe [7]. In certain situations, people could be even indifferent between an algorithm’s advice and a human’s advice, indicating that task objectivity might facilitate people’s acceptance of algorithms. However, Logg et al [20] found out that people always prefer decision input from algorithms, in both objective and subjective tasks, leaving uncertainty about whether task objectivity and subjectivity play a role or not.

Second, although both streams of research have a common goal of facilitating the acceptance of algorithmic decision support systems, suggestions as to how to reach this goal largely diverge. Studies of algorithm appreciation argue that to increase the use of algorithmic input, people should be made aware of the algorithmic nature of the input [20]. On the other hand, studies of algorithm aversion argue that designers should emphasize the “human touch” of the input to facilitate its acceptance [20]. With algorithmic decision aids increasingly used in industry, it is crucial to understand which approaches are more promising in facilitating adequate use.

Third, the two streams of research also lead to different trajectories for future research. While the stream of algorithm aversion tries to study how to alleviate the aversion effect by emphasizing the human aspects of algorithms, the stream of algorithm appreciation calls for more investigations in how we can further people’s reliance on algorithms by increasing the transparency in AI systems [20]. Seeing all these differences, we believe it is crucial to clarify the relationship between these opposite attitudes toward algorithms, i.e., are algorithm aversion and algorithm appreciation two distinct phenomena, each with its own cause, or can they be the two ends of one continuous spectrum sharing a single factor?

2.2 Reconciling Algorithm Aversion and Algorithm Appreciation

Current findings from research on algorithm aversion and algorithm appreciation are seemingly at odds with each other. We do not fully understand the mechanisms that drive the occurrence of one phenomenon over the other. A possible explanation is that, with time, people are getting more familiar with algorithms, and algorithms are also becoming more powerful, thus the transition from algorithm aversion to appreciation. This reasoning, however, does not explain well why both phenomena still co-exist in recent studies, indicating the existence of other factors. Following, we discuss studies with mixed results. Such studies might provide insights about how these two seemingly contradictory phenomena can be reconciled.

Longoni and Cian’s study of marketing [22] found that people prefer algorithmic over human recommendations if the goal is utilitarian and vice versa if the goal is hedonic. This finding not only echoes that of Castelo et al’s research on perceived subjectivity and objectivity but also has important implications because it shows that the distinction between algorithm aversion and algorithm appreciation is not absolute. While Castelo et al found an algorithm aversion effect (even if they made the task seem more objective, the best they can achieve was indifference between algorithms and humans and never can they get an algorithm appreciation effect), Longoni et al’s study shows that there might be a common factor with which we can “nudge” the effect toward the appreciation a bit and make the perceived subjectivity-objectivity the key factor deciding between algorithm aversion and algorithm appreciation.

So, what can be the “nudging” factor? A study by Bigman and Gray [3] seems to give some insights. They conducted nine different studies on people’s attitude toward moral judgements made by algorithms and other people. While most of the studies, across different contexts, showed an algorithm aversion effect, study 9 showed, surprisingly, an algorithm appreciation effect when the algorithm (a system called HealthComp) was presented as having a 95% success rate and the human (Dr. Jones) only had a 75% success rate. That is, the difference in the description indicated a clear difference in the competence, and this difference might have reversed the outcome that would otherwise be an algorithm aversion effect.

Other relevant studies also have discussed the role of perceived competence. Logg et al. mentioned that expertise might be an important factor in deciding between appreciation and aversion [20]. Thurman et al. [30] also noted that, in Logg et al’s study, the “human advice” came

from another experiment participant and not an expert, suggesting that this had shaped the results of the study. Both indicated the importance of how the advice giver is presented, i.e., the framing of the agents, in the study.

Along this line of discussion on the competence and the framing of the decision-support agents (human or algorithm), we look further in relevant studies that have compared humans versus algorithms regarding how each agent was framed. The comparison of several key studies is summarized in Table 1. Here we observe a systematic tendency: in general, how the algorithm and human conditions were described in a study seems to be related to its results. In studies that found an algorithm aversion effect, the human was often described as having high competence, such as doctors, physicians, experts, or “a very qualified person”, while in studies that found algorithmic appreciation, the description was “other people” or other participants in an experiment. This leads us to speculate whether this difference in competence framing is one key factor that determines how much people are willing to take in decision inputs, hence the distinction between algorithm aversion and algorithm appreciation. It is possible that when the human is framed as more competent than the algorithm, people will exhibit algorithm aversion behavior. In contrast, when humans are framed as less competent than the algorithm, we can observe an algorithm appreciation effect.

Table 1. Comparison of AI-Supported Decision Making Studies (Framing of Human vs. Algorithm)

Empirical Study	Part	Human Description	Algorithm Description	Result ^a
Castelo et al. (2019) [7]	Study 1	A very well qualified person	An algorithm	AVER
	Study 3	A qualified human	An algorithm	AVER
Promberger and Baron (2006) [25]		A physician	A computer program	AVER
Fuchs et al. (2016) [13]		A human expert	A software program	AVER
Longoni et al. (2019)[21]	Study 1	A physician	A computer	AVER
Önkal et al. (2009) [23]		A financial expert who makes good stock price predictions	A statistical model that makes good stock price predictions.	AVER
Bigman and Gray (2018) [3]	Study 1	A human driver	An autonomous computer program	AVER
	Study 2	A state committee consists of legal and mental health experts as well as representatives of the community.	CompNet, a super computer used by various government agencies for calculations, estimates, and decision-making.	AVER
	Study 3	Dr. Jones, a doctor with a great capacity for both rational thinking and for emotional compassion.	HealthComp, an autonomous statistics-based computer system with a great capacity for rational thinking, but totally lacking in emotional compassion.	AVER

Bigman and Gray (2018) [3]	Study 4	Colonel Jones, an officer with a great capacity for both rational thinking and for emotional compassion.	CompNet, an autonomous statistics-based computer system with a great capacity for rational thinking but is totally lacking in emotional compassion.	AVER
	Study 9	Dr. Jones, who had a 75% success rate	HealthComp, which had a 95% success rate	APPR
Longoni and Cian (2020) [22]	Study 1	A person	An algorithm	MIXED
Prahl and Van Swol (2017) [24]		Steven, a person experienced in operating room management issues.	An advanced computer system	MIXED
Logg et al. (2019) [20]	Study 1A	Participants from a past experiment	An algorithm ran calculations based on estimates of participants from a past study.	APPR
	Study 1B	An aggregation of 275 other participants	An algorithm	APPR
	Study 1C&1D	48 people in another study	An algorithm	APPR
	Study 2	Another participant	An algorithm	APPR
	Study 4	A randomly chosen participant from a pool of 314 participants who took a past study.	An algorithm, based on estimates of 314 participants who took a past study.	APPR
Dijkstra et al. (1998) [10]		A human	An expert system	APPR
Gunaratne et al. (2018) [16]		Crowdsourced: the average allocation made by other people in the study	Algorithmic: the recommended allocation based on recent research	APPR

^aAPPR: Algorithm Appreciation; AVER: Algorithm Aversion; MIXED: Mixed results or difference not significant

2.3 Summary and Research Questions

In sum, previous research suggests that the framing of the human and algorithmic decision-support agents plays a major role in eliciting an aversion or appreciation response. More specifically, we raise three broader research questions:

RQ1. How does the framing of the agents (the human and the algorithm) affect algorithm aversion and algorithm appreciation effects? That is, can we change an algorithm aversion situation into an algorithm appreciation one (or vice versa) by simply changing the framing of the agents?

Previous research found contradicting results regarding task type, specifically task objectivity and subjectivity, on people’s preference over algorithmic input versus human input. Some

research shows that when tasks are perceived as more objective, people are less averse about algorithms [7]. However, some other research shows that people always prefer algorithmic input regardless of task type [20], leaving a theoretical inconsistency that must be resolved. Therefore, while we are investigating the effect of framing in determining between algorithm aversion and algorithm appreciation, we also plan to investigate whether task type plays a role, and whether there is any interaction between task type and framing.

RQ2. Does task type play a role in determining between algorithm aversion and algorithm appreciation? Also, what is the relationship between task type and framing? Is there any interaction?

Lastly, to test whether the outcome of the difference in framing is actually related to, as we observed, the perceived competence, we leverage Raven et al's Interpersonal Power Inventory [26] and its root, the framework of power sources [11] from the field of Organizational Behavior. Power, or social power, is a fundamental concept in Organizational Behavior, usually defined as "the capacity to influence others" [33]. By this definition, the reason why people can influence others is because they have more power over other people. Note that this power is not just the narrowly defined power that leaders have over their subordinates, but in a much broader sense, encompassing many ways in which one can influence the others. In this regard, especially worthy of noting is French and Raven's framework, in which they specified six types of power, one of them being *expert power*, which is the type of power that some people holds because they have more knowledge or competence than others and therefore have influence over other people. We believe that this concept of influence is especially relevant in discussing how algorithmic and human decision-support agents, because of their perceived competence, influence people's decision making. Thus, we include this concept in the study to investigate whether our findings about framing, if any, are because of the perceived competence, i.e., expert power, or they are due to other types of factors in interpersonal influence.

RQ3. Is the perceived competence, i.e., expert power, of the human and the algorithm a crucial factor explaining the effect of different framings?

We explore the answers to these questions through three experiments. We plan to set the baseline with Experiment 1 and see if we can create an opposite effect in Experiment 2 by changing only the description of the agents, while everything else remain the same. We then demonstrate in Experiment 3 how the findings in the first two experiments can be used to explain inconsistent evidence regarding algorithm appreciation and algorithm aversion.

3 EXPERIMENT 1: THE BASELINE

In Experiment 1, we hope to set a benchmark for how much people are influenced by suggestions from humans versus algorithms. This benchmark can then serve as the baseline for the next study, where we can try to create an opposite effect.

3.1 Methods

3.1.1 Design. Experiment 1 leveraged a 2 (task type: creative vs. analytical) x 2 (agent type: algorithm vs. human) within-participants experiment design, resulting in four major blocks of questions which every participant went through in random order.

Following the definition of objective and subjective tasks in Castelo et al's study [7], the creative questions were subjective questions involving personal opinion. We have developed three similar creative questions, including "A and B are two abstract paintings. Among 100

general people, exactly how many do you think will find painting B more creative?”, “A and B are two descriptions for this image. Among 100 general people, exactly how many do you think will find B more creative?”, and “A and B are two drawings of a mosquito. Among 100 general people, exactly how many do you think will find drawing B more creative?”

The analytical questions, on the other hand, asked objective and quantifiable questions, including “There are one red line and one blue line in this picture. The sum of their length is 100 inches. Exactly how many inches do you think the red line is?”, “By mixing two colors with different ratios, we can create many in-between colors. Now, below is a color mixed from purple and red (these three colors were shown). What do you think is the exact percentage of purple?”, and “(After showing a graph for 10 seconds) The graph in the previous page has 100 shapes in total, including green triangles and orange dots. Exactly how many of them are green triangles?”

Three questions of the same type formed a question set as repeated measures, and the order of them was randomized for every participant. For the need of the experiment design, there were two sets of creative questions that were highly similar in their format and content. For each participant, one set of them was matched with an algorithmic suggestion provider to form a question block, and the other set was matched with a human suggestion provider to form another block. This configuration was the same for analytical questions, which also had two sets of questions matching randomly with the other algorithm and the other human, forming the other two blocks. The two human suggestion providers and the two algorithmic suggestion providers only differed in their names and descriptions.

3.1.2 Material and Procedure. The experiment and data collection were carried out using a Qualtrics survey. After accepting the task on Amazon Turk, participants were guided to Qualtrics, read the instructions, and gave their consents to participate in this study. They were then presented with the four question blocks in random order. The overall structure of a question block is shown in Fig. 1, and we will explain in detail in the next few paragraphs.

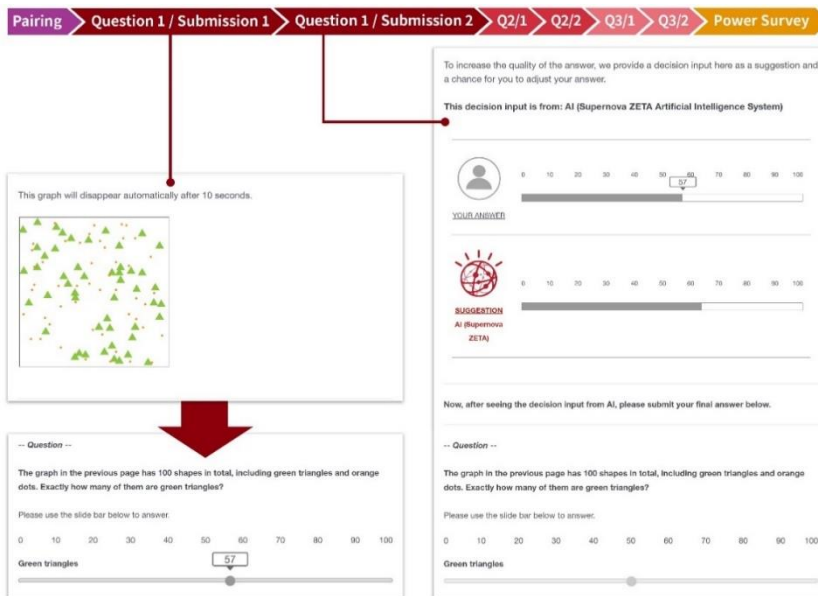


Fig. 1. The structure of a question block. Every participant went through four question blocks (corresponding to 2x2 conditions) in random order. This shown block is Analytical Task x Algorithm.

In each block, participants were firstly paired with an agent. There were totally four agents in experiment 1: *AI (Galaxy II Artificial Intelligence System)*, *AI (Supernova ZETA Artificial Intelligence System)*, *Another Mturker (AAX*****WY2K)*, and *Another Mturker (A99*****666PN)*. The main manipulation of agent type was in the descriptions of these agents, as shown in Table 2 in the section of Experiment 2.

The participants were then given the three creative or analytical questions in that block. All questions were written so that the answer was a number between 0 and 100. For every question, participants first read the question and submitted their initial answers with a slider. In the next page, they then saw the agent's "decision input", i.e., the suggestion, alongside their own answer. The decision input was actually calculated from their submitted answer in the previous page. It was always their initial answer plus or minus a random number between 6 and 9, regardless of the task type and the agent type (hence, the only difference between a human's suggestion and an algorithm's suggestion was how it was labeled). After seeing the suggestion, the participants then submitted their final answer. This two-stage submission came from the judge–advisor system paradigm [20,29]. In the original version of this paradigm, the decision input was always very close to the correct answer. We redesigned it this way so that the participant's attitude toward the agent was not affected by the distance between their initial answer and the agent's suggestion.

After the three questions, at the end of each block, participants were given a social power survey that measured their attitudes toward the agent, which will be explained in detail in the measures section.

The participants went through all four blocks of the same format (pairing with a suggestion provider → three questions → social power survey). They were then asked to provide personal information such as gender, age, and cultural background, and to answer one attention check question. Finally, they were debriefed about the experiment manipulation that these decision inputs were all calculated and there was no fundamental difference between the agents except that they were labeled differently.

3.1.3 Measures. We measured how much the agent's suggestion influenced the participant's by calculating, for every question, how much the participant changed from the initial answer to the final answer, divided by the difference between the initial answer and the decision input (which was always between 6 and 9 or -6 and -9). This value, which we called "influence factor", often ranged from 0 to 1 (0% to 100%). If the participants did not change their answers at all, it was a 0% influence. If their final answer was exactly as the suggestion, it was a 100% influence. Since we had three questions in each condition, the influence factor of that condition was defined as the mean of three individual influence factors.

The social power survey was adapted from Raven et al's Interpersonal Power Inventory [26], which was based on French and Raven's framework of power sources [11]. The original inventory followed this framework and measured all 6 types of social power sources. However, due to incompatibility with our experiment design, we deleted four parts that can only be answered after long-time interpersonal interaction. We therefore only used the remaining four parts that measured four sources (types) of power: Expert power, Referent power, Information power, and Legitimate power-position. For each type of power, there were three questions. Participants answered each with a 7-point Likert scale, ranging from Strongly Disagree to Strongly Agree, coded as 1 to 7 points. The participant's rating for a certain type of power was thus calculated by summing the answers of all three questions.

3.1.4 Participants. We recruited 120 participants on Amazon's Mechanical Turk (44 females, 75 males, and 1 prefer not to say) with an average age of 38.85. All these Mturkers met the criteria we set, that only those who were in the US and had a task approval rate of 90% or higher can participate. All participants were paid \$4.50 for participating in this study.

From a pilot study, we have learned that many participants did not answer the questions properly on Mechanical Turk. The problem was especially serious for the current experiment design, probably because of its complexity and its length. Therefore, before collecting data, we had defined the criteria of what can qualify as a usable data entry from one participant. First, the participants must pass the attention check question asking what questions and graphs they had seen in this study. Second, due to the experiment design, the participant's initial answers cannot be too high or too low to let the decision inputs be larger than 100 or smaller than 0, which cannot be displayed properly. Third, normally, we would expect a participant's final answer should be between their initial answer and the decision input, i.e., the influence factor should be between 0% and 100%. However, we observed in the pilot study that many participants were just dragging the sliders randomly. We therefore mandated that, for one participant's data entry to be valid, all of his or her influence factor should be between -50% and 150% (we added a 50% buffer zone in both directions to the original 0%-100% to accommodate for minor errors).

We learned from the pilot study that only around one thirds of the participants will pass all three criteria, so we recruited three times of the needed participant number (power analysis suggested that we only needed 40). These screening criteria were pre-registered on OSF before we started collecting data: https://osf.io/kh82r/?view_only=8c7a02b0848c44b69783f8101003ae76

3.2 Results

3.2.1 Preliminaries. Among the 120 participants that we recruited, 41 did not pass the attention check questions, 23 had answers that made the decision inputs larger than 100 or smaller than 0, and 62 had at least one influence factor larger than 150% or smaller than -50% (we stated the rationales behind these three criteria in the Methods section). In total, that left us with data from 47 participants that met all criteria (17 females, 29 males, and 1 prefer not to say. $\text{Average}_{\text{age}} = 39.36$). The ratio was in the expected range we had found in the pilot study.

3.2.2 Influence Factor. A repeated-measures ANOVA showed that there was a significant main effect of agent type on the influence factor, showing a clear algorithm appreciation effect. Participants were influenced more when the suggestion came from algorithms ($M=0.57$, $SD=0.30$) than when it came from humans ($M = 0.45$, $SD = 0.29$), $F(1, 46) = 13.79$, $p < .001$; see Fig. 2 on the next page. There was no main effect of task type, $F(1, 46) = .27$, $p = .61$, and there was no interaction effect, $p = .51$.

To be more careful, we additionally used a bootstrapped variant of ANOVA with ANOVA.boot function from lmbot package in R to confirm these results because the Shapiro-Wilk normality test showed that the distribution of influence factor was not normal ($p < .001$). With this new method, the three p-values in the previous paragraph became 0.002, 0.68, 0.63, respectively. The main conclusion is the same.

3.2.3 Power. Through a repeated-measures ANOVA, we found a significant main effect of agent type on power. The suggestion giver was perceived to have more expert power when it was an algorithm ($M = 15.4$, $SD = 4.36$) than when it was a human ($M = 12.4$, $SD = 4.19$), $F(1, 46) = 21.18$, $p < .001$; see Fig. 3 on the next page. Also, the suggestion provider had more legitimate power-position when it was an algorithm ($M = 12.1$, $SD = 4.07$) than when it was a human ($M = 10.4$, SD

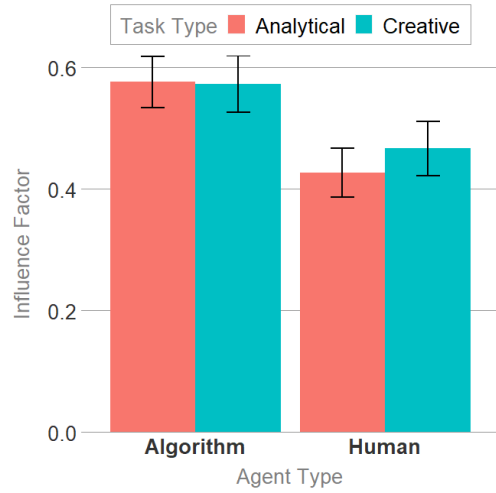


Fig. 2. Experiment 1: Mean Influence Factor by Task Type (Analytical vs. Creative) and Agent Type (Algorithm vs. Human) showing a clear algorithm appreciation effect. Error bars represent standard errors of the mean.

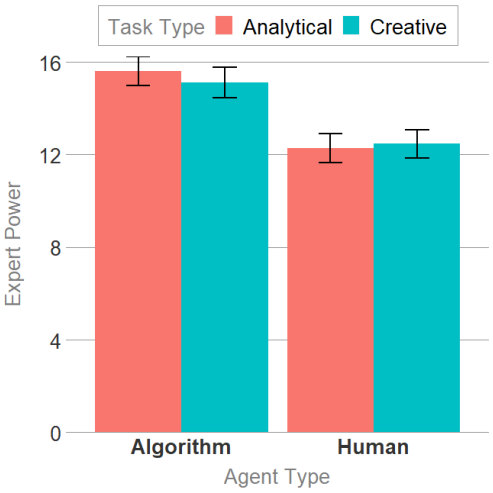


Fig. 3. Experiment 1: Total Score of Expert Power by Task Type (Analytical vs. Creative) and Agent Type (Algorithm vs. Human). The algorithm had higher expert power than the human. Error bars represent standard errors of the mean.

= 4.01), $F(1, 46) = 131.11$, $p < .001$. The same held true for informational power if it was an algorithm ($M = 15.3$, $SD = 3.72$) than when it was a human ($M = 14.0$, $SD = 3.66$), $F(1, 46) = 7.737$, $p = .008$. There was no significant difference in referent power, and there was neither main effect of task type, nor was there interaction effect between task type and agent type.

3.2.4 Mediation Analysis. To know whether power did mediate how much the participants were influenced by different suggestion providers, we also ran a mediation analysis. The result showed that the effect of agent type on influence factor was fully mediated via expert power. As Figure 4 illustrates, the regression coefficient between agent type and the influence factor and the

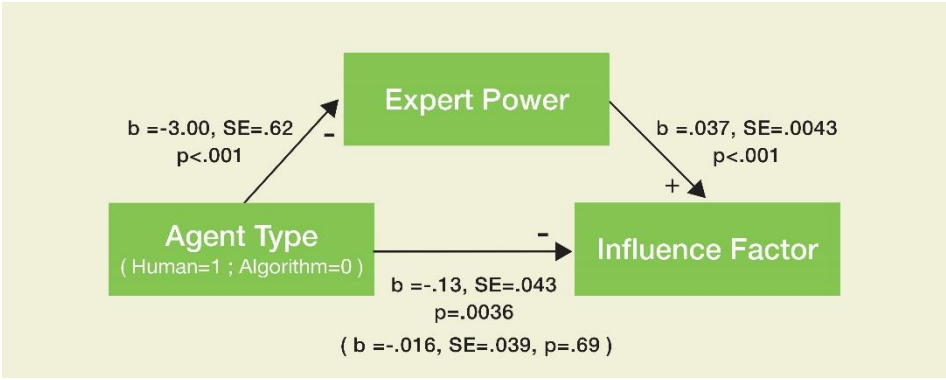


Fig. 4. Mediation analysis revealed that Expert Power fully mediated the effect of Agent Type on Influence Factor in Experiment 1. The indirect effect was significant.





regression coefficient between expert power and the influence factor were significant. The indirect effect was $(-3.0) * (.037) = -.11$. We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 1000

bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect was -.11, and the 95% confidence interval ranged from -.15 to -.07. Thus, the indirect effect was statistically significant ($p < .001$). When we included the mediators, expert power, in the regression, the effect of agent type on influence factor became insignificant, $b = -.016$, $SE = .039$, $p = .69$, which suggested that this effect was fully mediated by expert power. The other types of power sources, on the other hand, did not fully mediate the effect.

4 EXPERIMENT 2: THE REVERSION

In Experiment 2, we measure how much people are influenced by suggestions from humans versus algorithms. However, unlike Experiment 1, where results indicated an algorithm appreciation effect, in Experiment 2 we hope to show that we can create a different outcome if the framings of the algorithm and the human have been changed.

Table 2. Comparison of Experiment 1 and 2: The Framings of Agents and Results

	Experiment 1	Experiment 2
		
Human Description	Another Mturker ^(a) , who scored higher than most of the participants in a previous experiment.	A group of experts ^(c) , who formulated these decision inputs with their 20 years of experience in art education and creativity.
		
Algorithm Description	AI ^(b) , which scored higher than most of the participants in a previous experiment.	An algorithm ^(d) , which was created by aggregating several Mturkers' responses in a previous experiment.
Result	Algorithm Appreciation	Algorithm Aversion

^a "A99*****666PNJ" or "AAX*****WWYY2K" (Made-up Mturker Account Number)
^b "Galaxy II Artificial Intelligence System" or "Supernova ZETA Artificial Intelligence System"
^c "Group Da Vinci" or "Group Picasso" ^d "Mturkers Collection M" or "Mturkers Collection C"

4.1 Methods

4.1.1 *Design, Material and Procedure.* Experiment 2 was exactly as Experiment 1, with only one exception that the descriptions and the graphs of the human agent and the algorithm agent were changed for a different framing. In Experiment 1, the algorithm was perceived to have more expert power. Therefore, we tried to design the framing in Experiment 2 so that the human would appear to have more expert power here than in Experiment 1; conversely, we designed the

framing of the algorithm so that it would appear to have less expert power here than in Experiment 1. The comparison of the framings in Experiment 1 and 2 is summarized in Table 2. For the human condition, we used a group of experts instead of just one expert because we did not want to create the tension between numbers (one expert vs. many Mturkers). Except for the difference in framing shown in Table 2, all other aspects of the experiment design, including randomizations and data handling criteria, remained identical.

4.1.2 Participants. Same as Experiment 1, we recruited 120 participants on Amazon's Mechanical Turk (42 females, 77 males, and 1 prefer not to say) with an average age of 37.68. These Mturkers also met the criteria we set, as in Experiment 1. All participants were also paid \$4.50 for participating in this study.

4.2 Results

4.2.1 Preliminaries. Among the 120 participants that we recruited, 37 did not pass the attention check questions, 19 had answers that made the suggestions larger than 100 or smaller than 0, and 71 had at least one influence factor larger than 150% or smaller than -50%. In total, that left us with data from 36 participants that met all criteria (13 females and 23 males, $\text{Average}_{\text{age}} = 39.58$). The ratio was also within the expected range we had found in the pilot study and Experiment 1.

4.2.2 Influence Factor. A repeated-measures ANOVA showed that there was also a significant main effect of agent type on the influence factor. However, unlike Experiment 1, it was an algorithm aversion this time. Participants were influenced more when the suggestion came from the humans ($M = 0.58$, $SD = 0.30$) than when it came from the algorithm ($M = 0.45$, $SD = 0.27$), $F(1, 35) = 9.15$, $p = .0046$; see Fig. 5 on the next page. There was no main effect of task type, $F(1, 35) = 1.63$, $p = .21$; however, there was a statistically significant interaction between task type and agent type, $F(1, 35) = 8.335$, $p = .0066$. Pairwise comparisons, using paired t-test, showed that the difference between the human agent's influence and the algorithm's influence was significant when the task type is analytical ($t = -4.50$, $p < .001$), but not when the task type is creative ($t = .77$, $p = .45$).

Similar to Experiment 1, we additionally did a bootstrapped variant of ANOVA on influence factor to confirm these results. The three p-values in the previous paragraph became 0.007, 0.28, 0.058, respectively. The interaction effect became insignificant. To be prudent, we took this insignificance as our result because the bootstrapped version is theoretically more robust when the data is not normally distributed.

4.2.3 Power. Through repeated-measures ANOVA, we also found a significant main effect of the agent type on different power sources. The suggestion giver was perceived to have more expert power when it was human ($M = 16.6$, $SD = 4.01$) than when it was algorithm ($M = 13.2$, $SD = 4.81$), $F(1, 35) = 21.15$, $p < .001$; see Fig. 6 on the next page. Also, the human agent had more legitimate power ($M = 11.8$, $SD = 4.20$) than the algorithm ($M = 10.4$, $SD = 4.22$), $F(1, 35) = 10.6$, $p = .0025$. The same held true for informational power: the human agent ($M = 15.5$, $SD = 3.46$) had more power than the algorithm ($M = 14.1$, $SD = 3.70$), $F(1, 35) = 8.749$, $p = .0055$. Unlike Experiment 1, there was also a significant main effect of the agent type on referent power. The human agent was perceived to have more referent power ($M = 12.6$, $SD = 4.43$) than the algorithm ($M = 10.8$, $SD = 5.13$), $F(1, 35) = 8.177$, $p = .0071$. There was neither a main effect of task type, nor was there any interaction effect between task type and agent type on all types of power.

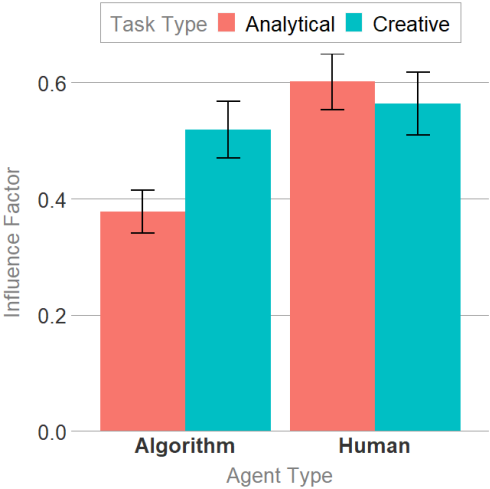


Fig. 5. Experiment 2: Mean Influence Factor by Task Type (Analytical vs. Creative) and Agent Type (Algorithm vs. Human) showing an algorithm aversion effect. Error bars represent standard errors of the mean.

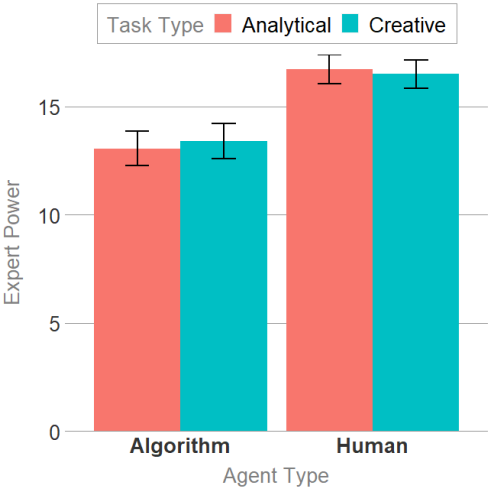


Fig. 6. Experiment 2: Total Score of Expert Power by Task Type (Analytical vs. Creative) and Agent Type (Algorithm vs. Human). The human had higher expert power than the algorithm. Error bars represent standard errors of the mean.

4.2.4 Mediation Analysis. We also ran a mediation analysis on expert power in Experiment 2. Again, the result showed that the effect of agent type on influence factor was fully mediated via expert power. As Fig. 7 illustrates, the regression coefficient between agent type and the influence factor and the regression coefficient between expert power and the influence factor were significant. The indirect effect was $(3.38) \times (.030) = .10$. We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 1000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect was .10, and the 95% confidence interval ranged from .055 to .15. Thus, the indirect effect was statistically significant ($p < .001$). When we included the mediators, expert power, in the

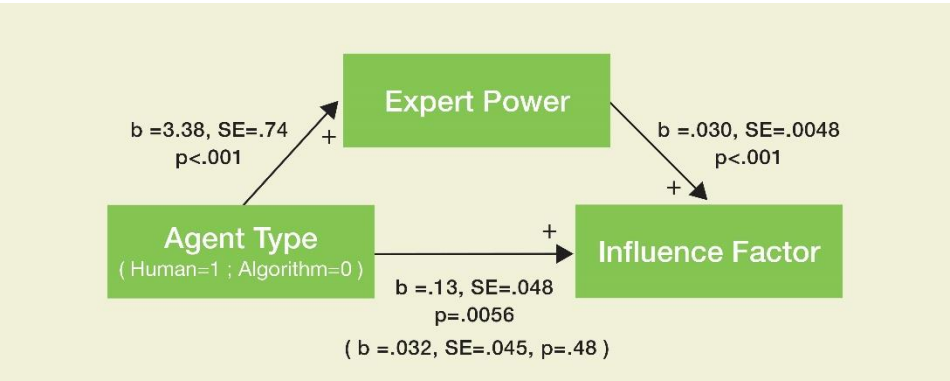


Fig. 7. Mediation analysis revealed that Expert Power fully mediated the effect of Agent Type on Influence Factor in Experiment 2. The indirect effect was significant.

regression, the effect of agent type on influence factor became insignificant, $b = .032$, $SE = .045$, $p = .48$, which, like Experiment 1 again, suggested that this effect was fully mediated by expert power.

4.2.5 Manipulation Check. To check whether the change of framings from Experiment 1 to Experiment 2 did influence participants' perception of expert power, we ran two t-tests to compare the agents' perceived expert power. Unpaired two-sample t-test revealed that there was a significant difference in expert power between human agents in Experiment 1 ($M = 12.37$, $SD = 4.19$) and human agents in Experiment 2 ($M = 16.61$, $SD = 4.01$), $t = -6.62$, $p < .001$. Human agents were perceived to have more expert power in Experiment 2 than in Experiment 1. On the other hand, unpaired two-sample t-test revealed that there was a significant difference in expert power between algorithmic agents in Experiment 1 ($M = 15.37$, $SD = 4.36$) and algorithmic agents in Experiment 2 ($M = 13.24$, $SD = 4.81$), $t = 2.95$, $p = .0037$. Algorithmic agents were perceived to have less expert power in Experiment 2 than in Experiment 1. These results indicated that our manipulation achieved desired effect.

5 EXPERIMENT 3: A BIGGER PICTURE

To illustrate how our results in experiment 1 and 2 might partly explain the inconsistency in previous literature on algorithm aversion and algorithm appreciation, we conduct Experiment 3, where we reveal how different combinations of framings can lead to different conclusions.

5.1 Methods

5.1.1 Design, Material and Procedure. The basic configuration of Experiment 3 was similar to Experiment 1 and 2, all involving the two-stage answer submission: for a given question, the participants submitted their initial answer first, and they then saw the suggestion from the agent and submitted their final answer. However, the design and the procedure of Experiment 3 were adjusted to address the need of this experiment. We dropped the manipulation of task type and focused more on the agent type and different framings.

Experiment 3 leveraged a 6-condition mixed design. Among the six conditions, three conditions were humans and the other three were algorithms, resulting in an underlying factor of agent type. The descriptions of conditions are summarized in Table 3 on the next page. In this study, each participant was randomly given one human condition and one algorithm condition. Each condition was paired with a set of analytical questions (each set included two questions, which were selected from Experiment 1 and 2) to form two main question blocks. Therefore, the manipulation of agent type was within-participants, while the manipulation of conditions was between-participants. The pairing of participants to conditions, the order of the conditions, and the matching between a condition and a question set were all randomized. For better screening, we also increased the difficulty of the attention check questions at the end of this study, asking the participants to precisely recognize the question they had answered. Besides these differences, the structure and procedure were the same as Experiment 1 and 2. Participants also went through the two question blocks following the same order (pairing with a suggestion provider → two questions → social power survey).

5.1.2 Participants. To accommodate a higher number of conditions and the transition from a within-participants design to a between-participants design, we recruited 408 participants on Amazon's Mechanical Turk (136 females, 271 males, and 1 prefer not to say). These Mturkers met

the criteria we set, that only those who were in the US and had a task approval rate of 90% or higher can participate. The average age was 36.04. All participants were paid \$3.00 for participating in this study.

Table 3. Conditions in Experiment 3

Agent Type	Condition	Description ^a
Algorithm	AI (Exp 1)	AI (Galaxy II Artificial Intelligence System), which scored higher than most of the participants in a previous experiment.
	Algorithm (Exp 2)	An algorithm (Mturkers Collection C), which was created by aggregating several Mturkers' responses in a previous experiment.
	Computer	A Computer, which is randomly picked from a previous experiment.
Human	Expert (Exp 2)	A group of experts (Group Da Vinci), who formulated these decision inputs with their 20 years of experience.
	Mturker (Exp 1)	Another Mturker (A93*****810PNJ), who scored higher than most of the participants in a previous experiment.
	Person	A person, who is randomly picked from a previous experiment.

^a The images of AI, Algorithm, Expert, Mturker were the same as Experiment 1 and 2. The image of Computer was the same as Algorithm, and the image of Person was the same as Mturker.

5.2 Results

5.2.1 Preliminaries. Among the 408 participants that we recruited, 230 did not pass the attention check questions, 44 had answers that made the suggestions larger than 100 or smaller than 0, and 180 had at least one influence factor larger than 150% or smaller than -50%. In total, that left us with data from 129 participants that met all criteria (43 females, 85 males, and 1 prefer not to say, Average_{age} = 35.76). Although we have increased the difficulty of the attention check questions, the overall pass ratio was similar compared to Experiment 1 and 2.

5.2.2 Influence Factor: Overall Analysis. We first tested the within-participants effect of agent type on influence factor. A paired t-test revealed that the difference between the influence of humans ($M = 0.52, SD = 0.36$) and the influence of algorithms ($M = 0.56, SD = 0.34$) was not significant, $t = 0.90, p = .37$, given our study design and this specific set of framings.

We then tested the effect of six different conditions on influence factor, see Fig 8 on page 18. One-way ANOVA showed a significant difference between six conditions, $F(5, 252) = 4.124, p = .0012$. Again, we did the bootstrapped version of ANOVA to check this result, $p = .0024$. A Tukey post hoc test revealed that there was a significant difference between Expert ($M = 0.69, SD = 0.32$) and Mturker ($M = 0.41, SD = 0.32$) and between Expert and Person ($M = 0.44, SD = 0.37$). The results are summarized in Table 4 (left) on page 17. With this study design, the variance between conditions was mainly because of the variance between different human agents.

5.2.3 Influence Factor: Unadjusted Pair Comparison. The main purpose of this experiment was to demonstrate how different combinations of framings can lead to different results. To do so, we additionally did unadjusted pairwise t-tests between conditions to simulate possible results if only two conditions were included in this study. The results showed that there were significant differences between AI ($M = 0.61, SD = 0.31$) and Mturker ($M = 0.41, SD = 0.32$), AI and Person (M

Table 4. Exp3: Comparisons Between Conditions, Including the Results of Multiple Comparison of Influence Factor (left), Pairwise Unadjusted Comparison of Influence Factor (middle), and Mediation Analysis of Whether Expert Power Mediated the Effect of Condition on Influence Factor (right)

Combination	Influence Factor: Multiple Comparison		Influence factor: Unadjusted Comparison (Results when a study only contains two conditions)		Expert Power: Mediation Analysis by Bootstrapping Method (Left: coded as 1; Right: coded as 0)			
	Tukey Test Adjusted <i>p</i>	<i>t</i> value	Unadjusted <i>p</i>	Possible Conclusion	Indirect Effect	SE	95% Confidence Interval	Mediation Occurred
AI - Algorithm	.74	-1.41	.16	-----	-.030	.021	[-.010, .074]	No
AI - Computer	.99	-0.67	.50	-----	-.032	.024	[-.013, .082]	No
Algorithm - Computer	.98	-0.66	.51	-----	-.0021	.022	[-.040, .045]	No
AI - Expert	.83	-1.30	.20	Inconclusive	-.021	.021	[-.067, .018]	No
AI - Mturker (Experiment 1)	.10	-2.76	.0071**	Algorithm Appreciation	-.075	.027	[-.027, .13]	Yes
AI - Person	.20	-2.31	.023*	Algorithm Appreciation	-.12	.033	[-.059, .19]	Yes
Algorithm - Expert (Experiment 2)	.10	-2.63	.010*	Algorithm Aversion	-.051	.022	[-.098, -.013]	Yes
Algorithm - Mturker	.80	-1.28	.20	Inconclusive	-.045	.024	[-.0003, .094]	No
Algorithm - Person	.94	-0.87	.39	Inconclusive	-.089	.029	[-.039, .15]	Yes
Computer - Expert	.42	-1.86	.066 +	(Marginal) Algorithm Aversion	-.053	.024	[-.11, -.011]	Yes
Computer - Mturker	.39	-1.91	.060 +	(Marginal) Algorithm Appreciation	-.043	.026	[-.0067, .096]	No
Computer - Person	.60	-1.50	.14	Inconclusive	-.087	.030	[-.033, .15]	Yes
Expert - Mturker	.0028**	-3.99	.00015***	-----	-.096	.028	[-.046, .15]	Yes
Expert - Person	.0063**	-3.52	.00068***	-----	-.14	.036	[-.074, .22]	Yes
Mturker - Person	.00	-0.40	.69	-----	-.044	.029	[-.0068, .11]	No

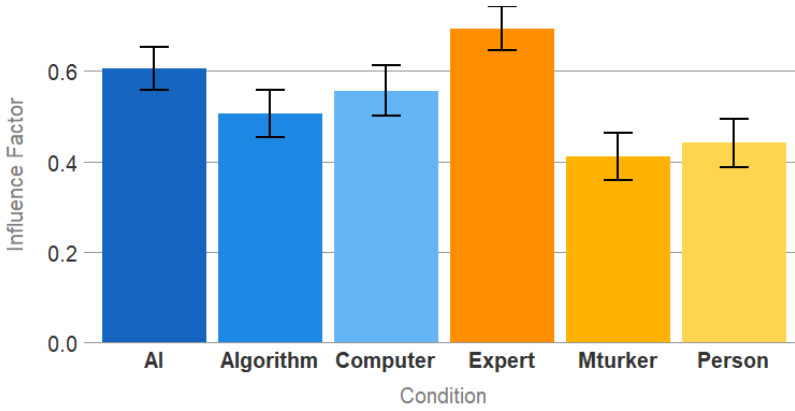


Fig. 8. Experiment 3: Mean Influence Factor by Condition. Error bars represent standard errors of the mean.

= 0.44, $SD = 0.37$), Algorithm ($M = 0.51$, $SD = 0.35$) and Expert ($M = 0.69$, $SD = 0.32$), Expert and Mturker, and Expert and Person. There were marginally significant differences between Computer ($M = 0.56$, $SD = 0.36$) and Expert, and Computer and Mturker. The results and possible conclusions are summarized in Table 4 (middle), showing that different combinations of framings can lead to different conclusions, which is likely a cause of inconsistent findings on algorithm appreciation and algorithm aversion in previous literature.

5.2.4 Power. We tested the within-participants effect of agent type on expert power. A paired t -test revealed that the difference between the expert power of humans ($M = 14.5$, $SD = 4.14$) and the expert power of algorithms ($M = 15.8$, $SD = 3.25$) was significant, $t = 2.65$, $p = .0090$. With our study design and this specific set of framings, algorithmic agents were perceived to have more expert power than humans overall.

We then tested the effect of six different conditions on expert power, see Fig 9. One-way ANOVA showed a significant difference between six conditions, $F(5, 252) = 10.58$, $p < .001$. A Tukey post hoc test revealed that there was a significant difference between AI ($M = 16.5$, $SD = 3.28$) and Mturker ($M = 14.00$, $SD = 3.89$), $p = .016$, AI and Person ($M = 12.6$, $SD = 4.14$), $p < .001$,

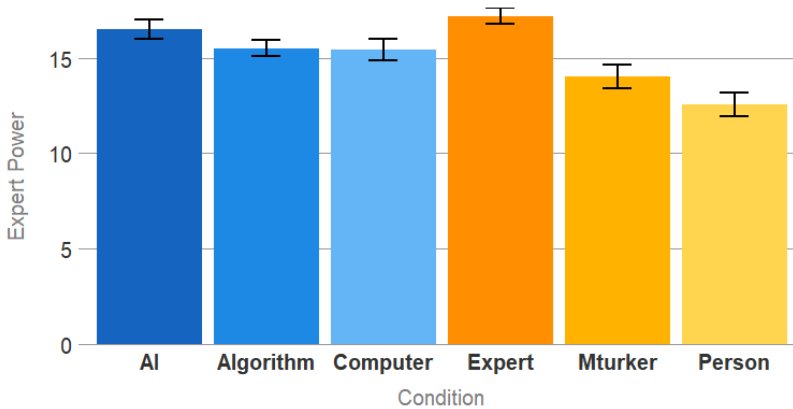


Fig. 9. Experiment 3: Total Score of Expert Power by Condition. Error bars represent standard errors of the mean.

Algorithm ($M = 15.5$, $SD = 2.89$) and Person, $p < .001$, Computer ($M = 15.4$, $SD = 3.53$) and Person, $p = .0016$, Expert ($M = 17.2$, $SD = 2.80$) and Mturker, $p < .001$, and Expert and Person, $p < .001$.

5.2.5 Mediation Analysis. Similar to the previous experiments, we were interested in whether expert power mediated the effects or not.

We first tested whether expert power mediated the effect of agent type on influence factor. Because the total effect of agent type on influence factor was not significant, we used the bootstrapping procedure to do mediation analysis instead of the traditional methodology that we used in Experiment 1 and 2. This was possible, though not ideal, because a significant total effect is not necessary to claim an indirect effect [27]. When human was coded as 1 and algorithm as 0, using Model 4 from Hayes's [17] PROCESS macro v3.5 beta for R, we calculated the indirect effect for each of 5000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect was $-.044$, $SE = 0.017$, 95% CI $[-0.079, -0.012]$. Because the confidence interval did not include zero, the indirect effect was significant. We would say that expert power significantly mediated the effect of agent type on influence factor.

Similarly, we also tested whether expert power mediated the effects of different conditions on influence factor. The bootstrapping method was basically the same, except for how conditions were coded: for each condition, we used five dummy variables to code the other five conditions and then calculated the indirect effect. The results are summarized in Table 4 (right). Although the results were mixed, we still can still see that expert power played an important role in the process. In all five pairs where the influence factor differed significantly, expert power significantly mediated the effects of conditions on influence factor.

6 DISCUSSION

6.1 Main Findings

Our results suggest that how we frame the algorithm and the human is a key factor deciding how much people are influenced by them, which in turn affects whether the experiment yields an algorithm appreciation result or an algorithm aversion one. In our first two experiments, we tested how the task type (creative vs. analytical) and the agent type (algorithm vs. human) affected how much participants were influenced in decision making tasks. In Experiment 1, participants were more influenced by the algorithm compared to the human, regardless of the task type, showing a clear effect of algorithm appreciation. In Experiment 2, we changed how the two agents were described and pictured in a way that increased the human's expert power and decreased the algorithm's expert power while all other factors remained unchanged. The result indicated that there was a significant main effect that, opposite to Experiment 1, participants now preferred the decision inputs from the human to those from the algorithm. By comparing Experiment 1 and Experiment 2, we can see that people's preference for algorithm vs. human is malleable through framing. We also showed that one key factor behind this malleability is expert power, a concept that we used to measure the agents' influence on the participants because of their perceived competence, which fully mediated the effect of agent type on the influence factor.

We further extended the idea of these findings in Experiment 3, comparing three types of framings of algorithmic agents and three types of framings of human agents. The results showed that different combinations of framings can lead to different conclusions about whether people prefer suggestions from algorithms or from humans, which explained the inconsistent findings

in previous literature. We also found that in pairs that differed significantly in influence factor, expert power was again the key variable that mediated the effect of different conditions on the influence factor. An interesting finding was that, given our framing design in Experiment 3, the variance of the influence factor between different human conditions was larger than the variance between algorithms. This is probably one important cause of the inconsistency in previous literature, and we believe that this finding provides an important insight and is worthy of further investigation in future study.

In sum, we found that the phenomena of algorithm appreciation and algorithm aversion have a common factor, expert power. In a given context, whether people will show algorithm aversion or algorithm appreciation depends on how powerful the algorithm is framed in comparison to the human. The key deciding factor between algorithm aversion and algorithm appreciation is therefore how much expert power each agent has *in relation to each other*. Thus, in our view, it is therefore less meaningful to argue whether people actually adhere more to algorithm inputs or human inputs. What matters more is the framing: *what kind of people* and *what kind of algorithm* we are comparing. And if the ultimate goals of these two streams of studies are both facilitating the application and acceptance of algorithmic decision support systems, the key question should become how we can better frame the algorithms so that they can be perceived as more competent, i.e., having more expert power, within a given context. We believe that, by shifting perspective in this way, we can generate fruitful findings of greater implication values.

6.2 Practical and Theoretical Implications

Our findings suggest that framing and expert power matter more than whether the decision input comes from algorithms or humans. We believe this finding is important in shaping the trajectory of future research in algorithmic decision support systems. Previously, the research stream that found algorithm aversion contradicted with the research stream that found algorithm appreciation in their suggestions to make these support systems more effective. The stream of algorithm aversion have addressed the importance of alleviating the effect of algorithm aversion and even have persuaded some companies using algorithms to present their outcomes as less algorithmic and with more “human touch” [20]. On the other hand, the stream of algorithm appreciation have suggested that, to make the decision inputs more influential, what we need to do is let people know that these inputs come from algorithms [20]. Our current study suggests that these two approaches are both not ideal. The key question is not whether the decision inputs are marked as from algorithms or from humans; instead, thinking about how these agents are framed and how to make these framings more powerful may generate more pertinent answers.

It should be noted that what we mean by “framing” is a broader term than how it is usually used. It consists of, traditionally, how a term is described and, additionally, how the term itself is chosen and the interaction between the term and the description. Take our Experiment 3 as an example. In the AI and Mturker conditions, their descriptions were both “who (which) scored higher than most of the participants in a previous experiment.” And in the Computer and Person conditions, the descriptions were both “who (which) is randomly picked from a previous experiment.” We would think, traditionally, that the framings of the algorithmic agent and the human agent were “the same” in both cases. However, with our simulation analysis in Experiment 3, we found an algorithm appreciation effect in the former pair, while it was inconclusive in the latter pair. It is therefore problematic to say that we can control the framing in a study by using the same description for both agents, and then claim that algorithm appreciation (in the former case) or no conclusion (in the latter case) is the default answer to whether people prefer

algorithms or humans—our results have shown that you can get different results even when the descriptions are held constant.

This example illustrates the complexity of framing in two ways. First, the term itself that one chooses also carries certain meaning and certain level of expert power, and this is inevitable when comparing algorithm agents against human agents. In this kind of study, we need to give different names to the agents so that people know that they are comparing different kinds of agents. However, by calling the human agent “another Mturker” instead of “a person”, and by calling the algorithm “AI” instead of just “a computer”, we already assign a certain amount of expert power to the agents, and this term becomes part of the framing, which influences people’s preference. Second, there might be interaction between the term and the description. The same description or even a single word might have different meanings according to the agent type. For example, “training” and “learning” might mean very different things to a human versus to an algorithm. “A group of experts” may sound powerful, but “a group of algorithms” is not that much more impressive than “an algorithm”. In short, when comparing humans and algorithms, framing and expert power are so entangled with agent type, the term we choose, the description, and the interaction between all these factors. We therefore should be very aware of these effects, and, although not entirely possible, try to control them carefully if we want to further investigate other factors in our future endeavors on this topic.

Our results also provide insights into some contradicting results in previous studies regarding task subjectivity and objectivity. In Castelo et al’s study [7], they found that the more objective the task is, the less severe algorithm aversion is. However, even in some of the most objective tasks, their participants were at most indifferent between the algorithmic input and the human input, indicating that algorithm aversion is overwhelmingly powerful. However, Logg et al’s study [20] suggested an overwhelming algorithm appreciation, regardless of task subjectivity and objectivity. It is therefore unclear regarding how task objectivity and subjectivity influence algorithm aversion and algorithm appreciation. Our findings reconcile this contradiction by showing that expert power is probably the overwhelmingly important factor here, while task type may be a minor influencer. It is possible that in Logg et al’s study, the power difference between human and algorithm was huge, so that they did not observe the effect of task type. Only when the power difference was smaller, such as in Castelo et al’s study, we can really see the smaller effect of task type.

This paper also contributes methodologically to the research on AI-supported decision making. Traditionally, research comparing the effect of algorithms and humans often let the participants choose between an algorithmic suggestion provider or a human suggestion provider [7,8,22] instead of actually measuring how much participants were influenced by these agents. The results, therefore, reflected the participant’s attitude toward algorithms versus humans but not the actual measurable behavior. To solve this problem, Logg et al. [20] adopted the judge–advisor system paradigm to measure influence when participants were given actual decision input in a decision making task. They also controlled the quality of the decision input by giving participants in different conditions identical advice, resulting in a clean experiment design where the only difference was whether the advice was labeled as coming from the algorithm or coming from the human. However, in their design, the advice was always a number very close to the actual answer, which raised a new issue. Since participants varied in their initial answers, some of them might had an initial answer close to the advice, while some might had an answer far from the advice. This difference might cause different participants to rate the quality of the advice differently, affecting their trust in the agents and their willingness to adhere to the advice. To address this

issue, we therefore refined this paradigm. Instead of giving the same advice to every participant, we gave participants decision input based on their initial answers. Since the decision input was always the participant's initial answer plus or minus 6 to 9, it was less viewed as too outrageous to follow. With this refined paradigm, we are therefore able to control both the quality of the decision input and participant's attitude toward the agent, which is an improvement to the original judge–advisor system paradigm.

6.3 Power as a Useful Framework in the Design of Decision Support Systems

A major finding of current study is the important role that expert power plays, which suggests great implication values of power-related theories in the design of decision support systems. Although the concept of power is traditionally applied only in human-human interaction, we think it is also applicable in human-algorithm interaction, especially for those algorithms that have human-like behavior or have replaced roles traditionally held by humans. As people are more likely to perceive these algorithms as autonomous agents, it is more likely that we can leverage the long traditions of Organizational Behavior and Social Psychology, where interpersonal interaction has been intensively studied.

In the current study, we have shown that by designing for more expert power, it is possible for an agent to cast more influence over people's decision making. It is worth noting that expert power is not the only type of power. According to French and Raven's framework [11], there are six types of power: Expert power, Reward power, Coercive power, Legitimate power, Referent power, and Information power. Expert power, which we are already familiar with, is the result of being more knowledgeable, competent, and knowing what the better action to take. On the other hand, people gain Reward power by controlling how much reward others can get, and they gain Coercive power when they can punish others who do not obey. Legitimate power is the authority and legitimacy given by organizations or social systems. Referent power, closely related to personal charisma, is held by people whose personality and personal traits attract admiration and identification from others. Finally, people who own Information power hold critical information at hand and can decide who can and cannot access that information.

Here we would like to propose that it is beneficial for us to incorporate the whole framework, including other types of power, which will lead to a broader view for the design of algorithms. Though this application of the power framework seems new, many parts of it have already been raised in previous research, just not under the umbrella of power framework. For example, Burton et al. have discussed ways to alleviate algorithm aversion [5]. They mentioned that lack of incentives to use algorithms inputs is one reason behind algorithm aversion, so they proposed solutions using economic and social incentivization. Seeing this suggestion through the lens of power framework, we would like to point out that providing economic incentive is a way to increase an agent's reward power. On the other hand, social incentivization, which is about creating a social context in which algorithms are trusted, is a way to increase an agent's legitimate power. Potentially, this framework not only covers such previous suggestions systematically, but also can serve as a convenient framework for brainstorming new ways to increase algorithms' influences: What if, along the line of legitimate power, we create a social context where some algorithms have authority? Or, along the line of referent power, can an algorithm become more influential if it has a good reputation? We believe this power framework can help us leverage knowledge from Organizational Behavior and Social Psychology and apply that on the design of decision support systems and, even more generally, the design of cooperative human-agent interaction.

6.4 Limitations and Future Directions

The current study is very much an initial investigation into reconciling algorithm aversion and algorithm appreciations, so it still leaves many open questions. First, we found that framing and expert power are crucial in deciding between algorithm aversion and algorithm appreciation, but it remains unclear how these factors work and interact to influence people's behavior. In Experiment 1, 2 and 3, we manipulated multiple dimensions in the framings. The descriptions differed in many ways: one vs. many, laymen vs. experts, "AI" vs. "algorithm" vs. "computer", how the algorithm is derived vs. its performance compared to general participants. We only had a vague sense about how these dimensions might work when designing these framings, and we still do not fully understand whether each dimension works or not, whether the effects of these dimensions would differ with different agent types, and how these dimensions might interact with each other. As previously discussed, when comparing algorithms and humans, the influence of framing is inevitable. It entangles with the terms and descriptions we use to refer to different types of agents. It is therefore difficult but crucial to have a deeper understanding of these dimensions that function as the building blocks of framing, how they work, and how they influence the perceived expert power. We therefore suggest that future research should investigate further how much these dimensions, including terms, descriptions, or even images, bear different levels of expert power, and how they affect people's attitude and behavior.

Second, the tasks used in this research were very limited. We only had one type of objective task (quantity judgement) and one type of subjective task (creativity judgement). It is therefore inconclusive whether the findings in this study can be generalizable to other types of tasks or other context, especially given algorithms' wide range of application that already exists today. The tasks used in this study all share certain characteristics. They are difficult to answer, yet they all seem to have an objectively correct answer (even the subjective questions such as "Among 100 general people, exactly how many will find painting A more creative?" should also have a correct answer through a large-scale survey study). This might make participants less confident in their own decisions and more likely to cling to the opinion of experts. It is very possible that people will trust their own idea more if the questions become more subjective, asking for their opinion instead of a correct answer. Besides, we also did not investigate tasks that involve social values such as moral judgement, or tasks that may lead to serious outcomes, like medical diagnosis. As the application of AI-supported decision making is growing in these scenarios, it will be pressing to test whether our findings can also be applied in these contexts. Also, with the design of this research, the relationship between agent type, task type, and expert power was highly simplified. In the real world, we would expect some level of interaction between them. For example, people are unlikely to believe an algorithm good at financial advice will also be good at reading X-ray images, but people might believe that an algorithm good at reading X-ray images might also be good at reading fMRI images (even if it is actually not). Future research is necessary to explore these relationships and interactions.

Third, while current research shows the general trend that people adhere more to the agent which has more expert power, there were still many participants whose performance did not align with this trend, indicating that there are more factors influencing people's use of decision support systems. Also, because of the limitation of online studies, we were not able to investigate in detail the rationale behind the perceived behavior. We believe there should be more factors underlying people's preference for (or aversion to) algorithms. This will continue to be a great challenge in understanding how we can design for algorithms' greater influence.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1942085 and Grant No. 1563705. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We wish to thank the anonymous reviewers for their feedback, and Chien Wen (Tina) Yuan and Johan Michalove for their valuable comments. We also want to thank Soyee Park and Wemi Oshun-Williams for their valuable input in the design of the tasks.

REFERENCES

- [1] Theo Araujo, Natali Helberger, Sanne Kruijemeier, and Claes H. de Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and Society* 35, 3: 611–623. <https://doi.org/10.1007/s00146-019-00931-w>
- [2] Benedikt Berger, Martin Adam, Alexander Ru, and Alexander Benlian. 2021. Watch Me Improve — Algorithm Aversion and Demonstrating the Ability to Learn. 63, 1: 55–68. <https://doi.org/10.1007/s12599-020-00678-5>
- [3] Yochanan E. Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181, March: 21–34. <https://doi.org/10.1016/j.cognition.2018.08.003>
- [4] Eric Bogert, Rick Watson, and Aaron Schecter. 2020. Algorithmic appreciation in creative tasks. *26th Americas Conference on Information Systems, AMCIS 2020*, July.
- [5] Jason W. Burton, Mari Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2: 220–239. <https://doi.org/10.1002/bdm.2155>
- [6] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. “Hello Ai”: Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW. <https://doi.org/10.1145/3359206>
- [7] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5: 809–825. <https://doi.org/10.1177/0022243719851788>
- [8] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 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>
- [9] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3: 1155–1170. <https://doi.org/10.1287/mnsc.2016.2643>
- [10] Jaap J. Dijkstra, Wim B.G. Liebrand, and Ellen Timminga. 1998. Persuasiveness of expert systems. *Behaviour and Information Technology* 17, 3: 155–163. <https://doi.org/10.1080/014492998119526>
- [11] Jr. John P. French and Betram Raven. 1960. The Bases of Social Power. *Group Dynamics*, January 1959.
- [12] Walter Frick. 2015. Here’s Why People Trust Human Judgment Over Algorithms. *Harvard Business Review*. Retrieved April 6, 2021 from <https://hbr.org/2015/02/heres-why-people-trust-human-judgment-over-algorithms>
- [13] Christoph Fuchs, Thomas Hess, Christian Matt, and Christian Hoerndlein. 2016. Human vs. Algorithmic recommendations in big data and the role of ambiguity. *AMCIS 2016: Surfing the IT Innovation Wave - 22nd Americas Conference on Information Systems*, August.
- [14] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW. <https://doi.org/10.1145/3359152>
- [15] Nina Grgic-Hlaca, Christoph Engel, and Krishna P. Gummadi. 2019. Human decision making with machine advice: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW. <https://doi.org/10.1145/3359280>
- [16] Junius Gunaratne, Lior Zalmanson, and Oded Nov. 2018. The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems* 35, 4: 1092–1120. <https://doi.org/10.1080/07421222.2018.1523534>
- [17] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.
- [18] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Twenty-Eighth European Conference on Information Systems (ECIS2020)*, June: 1–16.
- [19] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez I Badia. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2. <https://doi.org/10.1145/3415227>
- [20] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151, December 2018: 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>

- [21] Chiara Longoni, Andrea Bonezzi, and Carey K. Morewedge. 2019. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 46, 4: 629–650. <https://doi.org/10.1093/jcr/ucz013>
- [22] Chiara Longoni and Luca Cian. 2020. " Word-of-Machine " Effect: Shifts in Utilitarian and Hedonic Trade-offs " Word-of-Machine " Effect: Shifts in Utilitarian and Hedonic Trade-offs Determine Preference for (or Resistance to) Artificial Intelligence Recommenders. August.
- [23] Dilek Onkal, Paul Goodwin, Mary Thomson, Sinan Gonul, and Andrew Pollock. 2009. The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments. *Journal of Behavioral Decision Making* 22, February 2009: 390–409. <https://doi.org/10.1002/bdm.637>
- [24] Andrew Prah1 and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36, 6: 691–702. <https://doi.org/10.1002/for.2464>
- [25] Marianne Promberger and Jonathan Baron. 2006. Do Patients Trust Computers? *Journal of Behavioral Decision Making* 19, 2006: 455–468. <https://doi.org/10.1002/bdm.542>
- [26] H Raven. 1998. Conceptualizing and Measuring a PowerInteraction Model of Interpersonal Influence1. 307–332.
- [27] Derek D. Rucker, Kristopher J. Preacher, Zakary L. Tormala, and Richard E. Petty. 2011. Mediation Analysis in Social Psychology: Current Practices and New Recommendations. *Social and Personality Psychology Compass* 5, 6: 359–371. <https://doi.org/10.1111/j.1751-9004.2011.00355.x>
- [28] Donghee Shin, Bouziane Zaid, and Mohammed Ibrahine. 2020. Algorithm Appreciation: Algorithmic Performance, Developmental Processes, and User Interactions. *Proceedings of the 2020 IEEE International Conference on Communications, Computing, Cybersecurity, and Informatics, CCCI 2020*. <https://doi.org/10.1109/CCCI49893.2020.9256470>
- [29] Janet A. Snizek and Timothy Buckley. 1995. Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes* 62, 159–174. <https://doi.org/10.1006/obhd.1995.1040>
- [30] Neil Thurman, Judith Moeller, Natali Helberger, and Damian Trilling. 2019. My Friends, Editors, Algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism* 7, 4: 447–469. <https://doi.org/10.1080/21670811.2018.1493936>
- [31] Anna Williams, Imani Sherman, Simone Smarr, Brianna Posadas, and Juan E. Gilbert. 2019. *Human trust factors in image analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-94391-6_1
- [32] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32, 4: 403–414. <https://doi.org/10.1002/bdm.2118>
- [33] Gary A. Yukl. 2013. *Leadership in organizations*. Pearson.

Received April 2021; revised July 2021; accepted July 2021.