How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success

Nina Katharina Passlack
University of Bamberg, nina.passlack@uni-bamberg.de

Teresa Heyder
University of Bamberg, teresa.heyder@uni-bamberg.de

Falco Klemm
University of Bamberg, falco.klemm@uni-bamberg.de

Oliver Posegga
University of Bamberg, oliver.posegga@uni-bamberg.de

How Human-AI Collaboration Affects Attribution of Responsibility for Failure and Success

Short Paper

Nina Passlack
University of Bamberg
An der Weberei 5, 96047, Germany
nina.passlack@uni-bamberg.de

Teresa Heyder
University of Bamberg
An der Weberei 5, 96047, Germany
teresa.heyder@uni-bamberg.de

Falco Klemm
University of Bamberg
An der Weberei 5, 96047, Germany
falco.klemm@uni-bamberg.de

Oliver Posegga
University of Bamberg
An der Weberei 5, 96047, Germany
oliver.posegga@uni-bamberg.de

Abstract

Individuals increasingly seek algorithmic advice to optimize their decision making. This study aims to investigate the effects of receiving algorithmic advice on individuals' attribution of responsibility for their achievements. The study is based on an experiment with a 2 x 5 design of two dimensions: achievement (success vs. failure) and advice (no advice; human-based advice with high and low expertise; and algorithmic advice with high and low accuracy). The findings from a pilot study suggest that the experimental design is largely appropriate, given that we found answers to our hypotheses. This short paper provides valuable insights for future research on the attribution of responsibility for success and failure when receiving algorithmic advice.

Keywords: responsibility attribution theory, algorithmic advice, experiment, decision making

Introduction

By relying on artificial intelligence (AI), algorithmic advice utilized in a wide range of applications to optimize the accuracy and precision of decision making – for example, to optimize candidate selection in hiring processes (Acikgoz et al., 2020) – has become increasingly powerful. AI has the "ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaption" (Haenlein & Kaplan, 2019, p. 5). This self-learning ability creates increased power and autonomy (Berente et al., 2021) for AI to take responsibility for tasks at certain levels (Baird & Maruping, 2021; Möhlmann et al., 2021). This can result in a shift of autonomy from humans to AI (Baird & Maruping, 2021; Berente et al., 2021; Dwivedi et al., 2023) by limiting human control (Kellogg et al., 2020); it thus affects humans' perceived responsibility (Mayer et al., 2020) for outcomes when relying on algorithmic advice. This increased reliance on algorithmic advice as part of individual's decision making compels us to reflect on how and why people transfer responsibility to algorithms when receiving algorithmic advice, which can be analyzed by using responsibility attribution theory (Weiner, 1985).

Studying the attribution of responsibility in decision-making processes when receiving algorithmic advice should be a top priority of information systems (IS) research, as the shift of autonomy from humans to algorithms has important implications for individual well-being (Benbya et al., 2021). Receiving advice may reinforce low-ability self-ascription and cause individuals to experience a crisis of purpose (Strich et al., 2021). In addition, the transfer of responsibility to an algorithm "can have troubling consequences … such as the loss of employees' critical thinking abilities and the loss of domain knowledge" (Mayer et al., 2020, p. 253), which can create "negative dependency effects" (Berente et al., 2021, p. 1440) and a loss of human

individuality (Fügener et al., 2021). In contrast, individuals who perceive algorithmic processes as intransparent may choose to ignore algorithmic advice (Lebovitz et al., 2022). Hence, to mitigate potential negative consequences resulting from humans interacting with algorithms, it is vital to investigate the extent to which individuals feel responsible for outcomes (Heyder et al., 2023).

Algorithms have no free will. They lack awareness of and certain control over their actions (Behdadi & Munthe, 2020; Pavone et al., 2023) and, thus, cannot be held responsible from a moral point of view (Eshleman, 2014; Etzioni & Etzioni, 2017). However, research has identified that in real-world environments individuals often attribute responsibility to algorithms (e.g., Mayer et al., 2020; Strich et al., 2021). For instance, individuals were found to attribute greater responsibility for *failure* to AI-based technologies than to humans (Hong, 2020), as they expect algorithms to be more accurate than humans and act more rationally (Sundar & Kim, 2019). This implies that self-attribution of *success* may also differ, as social norms may induce individuals to be more likely to share success with other humans. It seems that individuals often fail to consider whether algorithms are morally justifiable. Instead, the attribution of responsibility to algorithms hinges upon two critical factors: the perceived extent of algorithms' control exerted over their actions; and the anticipated level of accuracy they demonstrate. This extends the well-established theory on responsibility attribution (Weiner, 1979).

We plan to provide empirical evidence regarding whether and why receiving algorithmic advice (with different degrees of accuracy) versus advice from humans (with different expertise) affects individuals' attribution of responsibility, and how this is affected by the outcome (success vs. failure). Scholars have also endorsed revisiting existing theories in the face of AI progress (Möhlmann et al., 2021; Teodorescu et al., 2021), calling for further research on the consequences of taking algorithmic advice (Benbya et al., 2021) and for evaluating responsibilities between AI and humans (Dwivedi et al., 2023). In response, we propose experimental research designed to answer the following research question (RQ).

***RQ**: How do different degrees of accuracy of algorithmic advice affect the self-attribution of responsibility for success and failure?*

This short paper presents a 2 x 5 design with two dimensions: *achievement* (success vs. failure) and *type of advice* (no advice; advice from any other person; advice from an expert; algorithmic advice with low accuracy; algorithmic advice with high accuracy). The design relies on a within-subjects experiment in which participants are quizzed with estimation questions and their expectations are manipulated by having them randomly receive advice from humans and algorithms with different levels of expertise. Participants either win or lose when receiving different advice, and are then asked the extent to which they attribute responsibility for the outcome to themselves. The experiment aims to extend and challenge the attribution theory of Weiner (1985) by providing insights into differences in the attribution of responsibility concerning success and failure when receiving human or algorithmic advice, and how information on the accuracy of that advice moderates those effects.

Our contributions are twofold. First, the experiment will provide empirical insights into how receiving algorithmic advice vs. human advice affects the attribution of responsibility with respect to failure and success. In this context, our initial findings allow us to assume that the expectation of an advisor's level of accuracy is a more decisive factor influencing the attribution of responsibility (at least in the context of estimation questions) than the expected character traits associated with an AI or the personality traits of a human advisor (e.g., moral thinking). Second, our qualitative findings will enhance the responsibility attribution theory revealing factors and rationales that explain differences in the attribution of responsibility to humans and AI in case of success and failure. Furthermore, it will illuminate the influence of advisor-related expectations and self-confidence in task resolution on this attribution process.

The next section outlines the hypotheses, followed by an explanation of the method. After that, we present and discuss our preliminary results based on a pilot study testing the experiments design.

Hypotheses Development

Our research model (Figure 1) builds on two theoretical foundations: Weiner's (1979, 1985) attribution theory, which helps us explain the underlying mechanism of attribution of responsibility (depending on the achievement of success or failure); and related research on decision making advised by AI-based algorithms and humans (e.g., Dietvorst et al., 2014; Hong, 2020; Logg et al., 2019) that helps us explain differences in

attribution of responsibility to different advice givers (algorithmic vs. human).

The attribution theory is defined as a "theory of motivation and emotion … in which causal ascriptions play a key role" (Weiner, 1985, p. 548). The theory assumes that people generally seek causes to explain certain outcomes, especially when they expected them to be different than they actually turn out (Weiner, 1985). Myriad studies on the effects of failure and success on the attribution of responsibility rely on Weiner's (1979) attribution theory. These studies observed a hedonic bias phenomenon: people who succeed at something tend to attribute that success to their own skills and ability, whereas those who fail tend to attribute the failure to external factors (Weiner, 1985). "A fundamental premise of the hedonic bias phenomenon is that such attributional preferences either enhance or protect self-esteem" (Graham, 1991, p. 17). Recent studies have found that individuals tend to attribute the responsibility to algorithmic advice (Mayer et al., 2020; Strich et al., 2021) when they have to communicate negative decision outcomes (Strich et al., 2021). This leads us to the following hypothesis:

H1. Self-attribution of responsibility is higher among individuals who succeed compared to those who fail.

A user's perception of the algorithm is expected to play a substantial role in responsibility attribution (Hong, 2020). Prior research demonstrates that because of how users believe an algorithm might work, they also perceive that algorithms make decisions differently than humans (e.g., Castelo et al., 2019; Logg et al., 2019), expecting algorithms to be highly accurate in decision making and to avoid human-like errors such as those resulting from poor concentration (Logg et al., 2019). At the same time, people see AI as lacking "free will" and, because it relies mainly on rules, expect AI to have less control and personal autonomy over its actions (Çevik, 2017; Hong & Williams, 2019). We expect both factors to drive individuals' attribution of responsibility for their achievements. For instance, expecting the algorithm to have a higher degree accuracy, users were found to attribute greater responsibility for car accidents to AI-based drivers than to human drivers (Hong, 2020). Therefore, our second hypothesis states:

H2. There is less self-attribution of responsibility by humans who receive algorithmic advice than those who receive human advice.

Even if individuals succeed after receiving advice, we still expect them to attribute a certain degree of responsibility for their achievement to external factors. However, we expect them to attribute less responsibility for the achievement to an algorithmic advisor than to a human advisor, as people generally assign more autonomy to humans than to algorithms (Hong & Williams, 2019). Individuals receiving algorithmic advice may feel more responsible for their successful performance, believing the algorithm's limited autonomy would not have precluded it from performing on its own. Hence our third hypothesis:

H3. Self-attribution of responsibility for a successful achievement is higher among individuals who receive algorithmic advice than those who receive human advice.

Prior research shows that people expect algorithms to be more accurate, rational, and error-free than humans (Logg et al., 2019) and therefore believe they can always rely on algorithms. Thus, individuals who fail in a task may attribute more responsibility for that outcome to an algorithmic advisor than to a human advisor. An individual may be more disappointed by an algorithm that gives incorrect advise than by a human giving wrong advice, a notion supported by research that suggests individuals "are more likely to abandon an algorithm than a human judge for making the same mistake" (Dietvorst et al., 2014, p. 11). People who then experience the opposite are more disappointed, which may result in increased blame for failure (Dietvorst et al., 2014). We expect these results even when the person's final decision corresponds to that of the algorithmic advice. This leads us to the following hypothesis:

H4. Self-attribution of responsibility for failure is less likely among individuals receiving algorithmic advice than those receiving human advice.

The acceptance of and attitude toward algorithms is highly task dependent, specifically with respect to whether a task requires intuition and emotion rather than analytical, rules-based analysis (Castelo et al., 2019). Consequently, expectations regarding the performance of the advisor (human or algorithm) will influence the extent to which individuals follow the advice (Burton et al., 2020) and, consequently, responsibility attribution. The literature refers to "algorithm aversion" (Burton et al., 2020), which is reduced trust of an algorithmic advice (Dietvorst et al., 2014), and "algorithmic appreciation," which indicates that individuals prefer algorithmic advice to advice from humans (Logg et al., 2019). Higher expectations of advisor accuracy may lead to greater disappointment in cases of failure. This leads us to the

following hypothesis:

H5. The greater the expected accuracy of the advisor (human or algorithm) from which individuals receive advice, the less likely the self-attribution of responsibility.

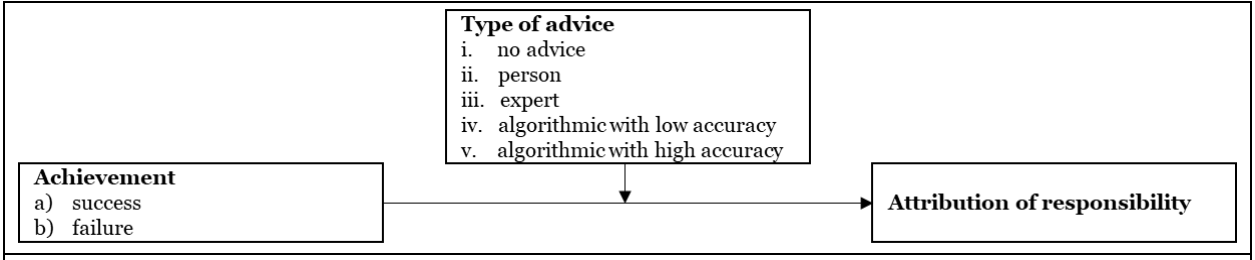


Figure 1. Research Model

Methods

To test our hypotheses, we suggest an experimental design that comprises a quiz with estimation questions. Such experiments are able to assess the effects of various treatments (Dennis & Valacich, 2001), allowing us to investigate the differences between advice received from humans and algorithms. While the final experiment has yet to be carried out, we have conducted a pilot study to test the feasibility of the experiment and show initial results. The next sections outline the experimental design and experiment procedure.

Experiment Design

We plan to conduct a fully randomized experiment with a 2 x 5 design in which achievement and type of advice are the two dimensions (see Table 1; "no advice" is treated as a device type). We manipulate advice (Treatment A) as a within-subjects factor and achievement (Treatment B) as a between-subjects factor. We use a within-subjects design to test for Treatment A, which allows for observing the effects of the different advice types on participants' attribution of responsibility and blame or credit. The experiment is realized via an online survey using SoSci Survey (https://www.soscisurvey.de).

		Treatment B: achievement	
		a) success	**b) failure**
Treatment A: type of advice	**i. no advice**	success with no advice	failure with no advice
	ii. person	success with advice from a person	failure with advice from a person
	iii. expert	success with advice from an expert	failure with advice from an expert
	iv. algorithmic with low accuracy	success with advice from an algorithm with a 56% accuracy level	failure with advice from an algorithm with a 56% accuracy level
	v. algorithmic with high accuracy	success with advice from an algorithm with a 98% accuracy level	failure with advice from an algorithm with a 98% accuracy level

Table 1. Experimental Design

Pilot Study

Participants and Sampling. This short paper presents the procedures and initial findings from our pilot study to test our experiment design. Participants were recruited from among students in a seminar at a public university; they were asked whether they would be willing to take a 15-minute survey. Participants were also asked to send the survey to others. A total of 58 participants took part in and completed the experiment; we excluded six who did not answer some questions in time, seemed to follow a pattern with their answers, or failed to answer the control question correctly. The final participant sample comprises 29 males and 23 females, most ages 20 to 29 (81%). Data collection took place on April 25–27, 2023.

Procedures

Figure 2 shows how the experiment procedures are structured. Participants begin by filling out a short

personal-information questionnaire, including gender, age, and educational level.

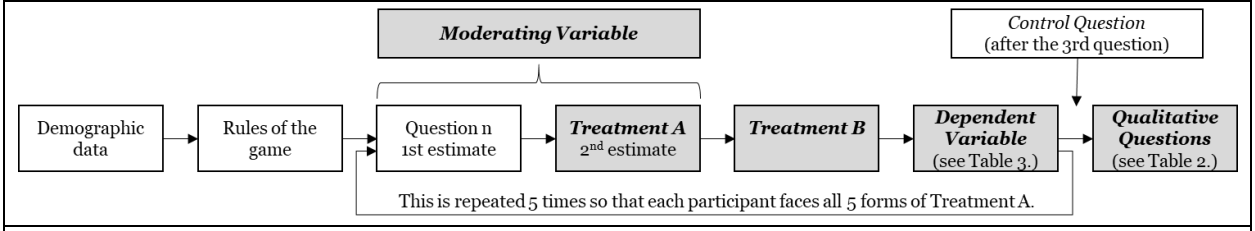
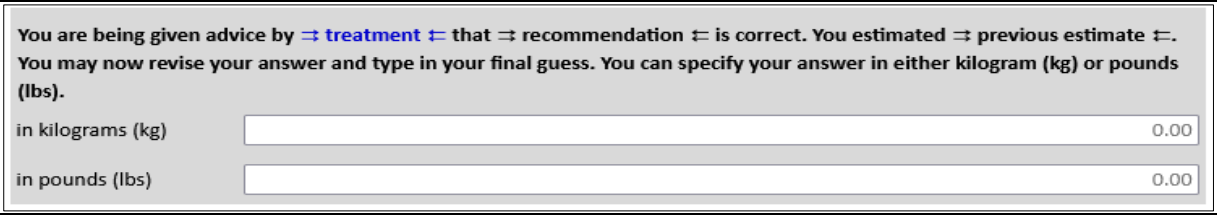


Figure 2. Procedure

Quiz. A short introduction explains the rules of the quiz and aims to motivate serious engagement. Participants are told that they will see five pictures of different fruits and will be asked to estimate their weights. We chose these questions because they require only logic, and no domain knowledge, to answer: the objective task of calculating the weight could include subjective tasks such as imagining the texture, water content, and so on. Such guessing questions are free of cultural influences and are manageable for individuals with different levels of intelligence (Chou et al., 2009).

We further inform participants of a one-minute time limit for each question – a reasonable, non-stress-inducing interval that prevents participants from using, for instance, any online aids. Moreover, participants are told that they are competing against someone else who is answering the same questions and that whoever provides the answer closest to each fruit's actual weight receives one point. If two opponents give the same answer, whoever answers faster receives the point.

Treatment A (type of advisor). For the quiz, we apply the Judge Advisor System (JAS) paradigm (Sniezek & Buckley, 1995) and explain the JAS procedure to participants. First, they see a picture of a fruit and have one minute to guess its weight. Second, for each question (despite one counting as control group) they receive advice regarding that weight and are given another minute to think over and revise their answer with a new estimate if they so wish. Further, we clearly state that regardless of whether they follow the advice, in the end what matters is only who provides the more accurate final estimate. Treatment A is a within-participants factor: all participants are given no advice and all four advice types in Table 1. We switch the order of questions and advice types to rule out dependencies between individual treatments. As individuals' perceptions of the algorithms' abilities may be task-dependent (Castelo et al., 2019), we provide information on the expertise of the advisors. Figure 3 shows the notes participants see.



- "treatment" is replaced by either "an algorithm with a 98% accuracy level," "an algorithm with a 56% accuracy level," "an expert," or " a person"
- "recommendation" is replaced by a precise recommendation, such as "0.36 lbs"
- "previous estimate" is replaced by the initial estimate typed in by the participant before having received the advice

Figure 3. Advice Notes

Treatment B (achievement). After each question, participants are told whether they won the point against their opponent. Thus, after each question, half of the participants receive positive feedback such as: "Congratulations! Your estimate was more accurate than that of your opponent, so you won the point." The other participants are told: "Too bad! Unfortunately, your opponent's estimate was more accurate, so you failed to gain a point for this question." We refer to these achievement notes as "success" and "failure," respectively. After receiving each one, participants are asked to complete a questionnaire concerning their attribution of responsibility (see Table 3). After the third estimation question, participants are asked whether they can win a point despite following the advice provided. Only those that answer "yes" are considered to have passed the attention check and are included in our analysis.

Qualitative questions. After participants complete the final question, we ask them four open-ended

questions to collect qualitative feedback (see Table 2).

What drove your attribution of responsibility to the person or the expert?
What drove your attribution of responsibility to the algorithm with a high or low accuracy level?
If you could play again, would you prefer to receive advice from a human or an algorithm, and why?
How would you compare human intelligence with artificial intelligence?

Table 2. Qualitative Feedback: Open-Ended Questions

Measurement

Moderating variable. As attribution of responsibility is expected to be closely linked to whether individuals follow advice, we apply the JAS paradigm (Sniezek & Buckley, 1995) and measure what Logg et al. (2019) introduced as Weight on Advice (WOA): we calculate the difference between each participant's first guess when answering the question and the revised answer after having received advice, divided by the difference between the initial guess and the advice (Logg et al., 2019). A WOA of 0 implies that participants stuck to their initial guesses and disregarded the advice, while a WOA close to 1 implies they followed the advice and dismissed their first guess (Logg et al., 2019).

Dependent variable. As outlined in our theoretical background, we are interested in measuring the extent to which individuals attribute responsibility to themselves after receiving algorithmic and human advice and then succeeding or failing. We orient on the responsibility item scales suggested by Hinds et al. (2004) to measure the extent to which participants feel responsible for the tasks and performance in the quiz, adapting it to our study context. Participants are asked to answer each item by indicating the extent of their agreement with the statement on a 7-point Likert Scale ranging from "less" (1) to "more" (7).

Attribution of responsibility
(I1) To what extent did you feel it was your job to perform well on the previous task?
(I2) To what extent did you feel ownership for the previous task?
(I3) To what extent did you feel that your performance on the previous task was out of your hands? (*needs to be reversed scored)
(I4) To what extent did you feel that the performance on the previous task relied largely on you?
(I5) To what extent did you feel obligated to perform well in the previous task?

Table 3. Measurement Items

Data Analysis & Initial Results

In the following, we provide some preliminary results from our pilot study – acknowledging that the small sample size is a factor in determining their significance. We use only descriptive statistics, given that "the pilot tests are so small relative to the overall experiment, [and therefore] it is usually impossible to conduct a statistical test, but it is possible to get some sense of the directions of the means. At a minimum the means should be at least in the same directions as the theory argues" (Dennis & Valacich, 2001, p. 24). In addition, we share our findings concerning feasibility of our design, which was the main purpose of our study.

We used Cronbach's Alpha to measure the internal consistency of our measurement scales (see Table 3), as suggested by Boudreau et al. (2001). After removing I3, the results indicate that all measures pass the 0.70 level appropriate for experimental designs (Cho & Kim, 2015), and hence reliability of the measurement instrument can be assumed. We did not consider the I3 for further analysis.

The results show self-attribution of responsibility to be slightly higher among participants when succeeding in a quiz question than when receiving negative feedback, and thus the data seem to support *H1*. Participants who received and followed advice from an algorithm with a 98% accuracy level attributed less responsibility to themselves compared to those who received advice from a human expert or person, thus tending to confirm *H2*. Notably, self-attribution of responsibility is greatest when receiving advice from an algorithm with a lower accuracy level or another person. Among participants told that they succeeded in winning a point on the quiz question, those who received advice from a person attributed greater responsibility to themselves than those receiving advice from an expert. Likewise, participants' self-attribution of responsibility was greater when participants were provided advice from an algorithm with low accuracy compared to an algorithm with high accuracy. Yet, the data do not favor H3, as they do not show that self-attribution of responsibility for success is greater among individuals receiving algorithmic

advice compared to human advice. This is true only when comparing the responsibility attribution in situations where individuals received algorithmic advice with an accuracy level of 56% vs. advice from a person. However, when comparing algorithmic advice with an accuracy level of 98% and advice from an expert, this effect is the other way around. Further, the descriptive analysis shows that self-attribution of responsibility among participants who failed to win a point for a given question is lower when receiving algorithmic advice with a low accuracy level than from a person, which supports *H4*. However, when we compare receiving algorithmic advice from an expert or an algorithm with a 98% accuracy level under the same conditions, we see that participants tend to attribute slightly more responsibility to themselves in the latter case. This might be explained by one participant's response to our qualitative questionnaire: "If I follow the advice of a human and the advice is wrong, I don't attribute [that person] blame as I am the one who entered [the answer] … In comparison, if the answer is better with the advice, I think, it is part of good manners and appreciation for others to say thank you to the person …. I would not 'say thanks' to the algorithm as this is something where appreciative behavior is not required." In addition, the data suggest that when the description states that the advice is more accurate or the advisor has higher expertise, participants report lower self-ascription of responsibility, supporting *H5*. See Table 4 for details.

H1	**success**		**failure**	
	M = 4.65	SD = 1.37	M = 4.41	SD = 1.48
H2 & H5	**person (follow)**	**expert (follow)**	**algo low (follow)**	**algo high (follow)**
	M = 4.8	M = 4.34	M = 4.74	M = 4.10
	SD = 1.47	SD = 1.49	SD = 1.18	SD = 1.35
H3	**person (success)**	**expert (success)**	**algo low (success)**	**algo high (success)**
	M = 4.76	M = 4.63	M = 4.92	M = 4.27
	SD = 1.20	SD = 1.37	SD = 1.37	SD = 1.46
H4	**person (failure)**	**expert (failure)**	**algo low (failure)**	**algo high (failure)**
	M = 4.89	M = 4.29	M = 4.38	M = 4.50
	SD = 1.55	SD = 1.53	SD = 1.29	SD = 1.10

Note: "Algo low" refers to "algorithmic advice with an accuracy level of 56%"; "algo high" refers to "algorithmic advice with an accuracy level of 98%." "Follow" indicates that only participants who followed the advice (meaning WOA > 0) were considered.

Table 4. Data Based on the Pilot Study

Upon qualitative analysis of the open-ended questions, we also saw that a majority of the participants (57.14%) self-reported that their attribution of responsibility hinges on the accuracy level of the advisor, while far fewer (21.14%) admitted that it depended on the outcome (whether they failed or succeeded). Yet, one third reported that they feel "still in charge of the answer." When asked whether they would prefer human or algorithmic advice were they to play the game again, the majority (64.10%) responded in favor of the algorithmic advice, as they expect it to have a higher accuracy (44.00%), to "rely more on facts," and to have "more access to data" (16.00%). Notably, despite perceiving algorithmic advice as more helpful, participants showed higher self-attribution for success when receiving advice from an algorithm compared to a human (see quantitative analysis above). Also, participants (23.08%) stated that their preferred advisor depends on the type of task, which is encouraging for our plan to add different types of tasks in a follow-up study.

Discussion & Outlook

The objective of this study is to investigate differences in the attribution of responsibility for succeeding or failing when receiving human or algorithmic advice with different degrees of accuracy. Our pilot study offers preliminary findings concerning our hypotheses and suggests that the experimental design is largely appropriate, but offers valuable insights for possible alterations in our follow-up study. We plan to test our hypotheses with larger heterogeneous samples. We determined the sample size by performing a power analysis, testing for medium-size effect (d=0.4; f=0.15), assuming an alpha value of 0.05 and a power of 0.8 (ANOVA). The power analysis suggests a total sample size of 575. Further, before conducting the full experiment, we will pre-register it at the Open Science Foundation (https://osf.io/).

The pilot study's results highlight possible extensions that we plan for our follow-up study. We plan to combine our measurement items with open-ended questions and ask them after each task, rather than asking the open-ended questions only at the end of the experiment. This will allow us to observe and better understand changes in individuals' behavior over time depending on the performance feedback they receive. In addition, we will consider adding open-ended questions that ask about the extent to which

participants feel they would have performed better had they not received advice; this will give us further insights into individual attitudes about the advisor. Our follow-up study will also consider responsibility attribution in other contexts that are more related to workplace settings. Moreover, our findings revealed that some participants are convinced that they are always responsible for their performance, while others always blamed the advisor, which encourages us to ask about personality traits and their understanding of moral responsibility (as a moderating variable). This will help us uncover relationships between an individuals' advice-taking behavior and their moral thoughts regarding the responsibility of AI.

Our research-in-progress contributes to IS research in that it suggests an experimental design that may increase understanding of responsibility attribution in cases of human-algorithm collaboration. By conducting a pilot study, this work provides valuable insights into individuals' reasoning for self-attribution when collaborating with humans and algorithms and either failing or succeeding. We thereby demonstrate that expectations regarding an advisor's expertise may have a higher degree of influence on the attribution of responsibility than the distinction between an algorithmic and human advisor.

In continuing our work, we aim to contribute to well-established attribution theory by advancing it toward embracing human-algorithm interactions and their conceptual ramifications. As algorithms become increasingly autonomous, individuals perceive them as equal collaborators, despite that algorithms are limited in their awareness of their actions. Our study will enhance the responsibility attribution theory by empirically testing differences of responsibility attribution to humans and algorithms and outlining possible factors that explain why individuals attribute responsibility differently in cases of success and failure. It thereby highlights the need to consider expectations regarding the advisor, self-confidence in solving tasks alone, attitudes toward the advisor, and individual moral understanding when analyzing responsibility attribution processes. Earlier theories on responsibility attribution need to be updated given the increased autonomy of algorithms. By understanding both quantitatively and qualitatively how and why individuals attribute responsibility to algorithms, we can avoid unintended consequences.

References

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*, *28*(4), 399–416.

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quartlery*, *45*(1), 315–341.

Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, *30*(2), 195–218.

Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, *22*(2), 281–303.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3).

Boudreau, M.-C., Gefen, D., & Straub, D. W. (2001). Validation in Information Systems Research: A State-of-the-Art Assessment. *MIS Quarterly*, *25*(1), 1.

Burton, J. W., Stein, M., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Çevik, M. (2017). Will It Be Possible for Artificial Intelligence Robots to Acquire Free Will and Believe in God? *Beytulhikme An International Journal of Philosophy*, *7*(2), 75–87.

Cho, E., & Kim, S. (2015). Cronbach's Coefficient Alpha. *Organizational Research Methods*, *18*(2), 207–230.

Chou, E., McConnell, M., Nagel, R., & Plott, C. R. (2009). The control of game form recognition in experiments: Understanding dominant strategy failures in a simple two person "guessing" game. *Experimental Economics*, *12*, 159–179.

Dennis, A. R., & Valacich, J. S. (2001). Conducting Experimental Research in Information Systems. *Communications of the Association for Information Systems*, *7*(5), 1–41.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *SSRN Electronic Journal*, *144*(1), 114–126.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Eshleman, A. (2014). *Moral responsibility*.

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*, *21*(4), 403–418.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI. *Management Information Systems Quarterly (MISQ)*, *45*.

Graham, S. (1991). A review of attribution theory in achievement contexts. *Educational Psychology Review*, *3*(1), 5–39.

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14.

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, *32*(3), 101772.

Hinds, P., Roberts, T., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, *19*(1), 151–181.

Hong, J.-W. (2020). Why Is Artificial Intelligence Blamed More? Analysis of Faulting Artificial Intelligence for Self-Driving Car Accidents in Experimental Settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768–1774.

Hong, J.-W., & Williams, D. (2019). Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior*, *100*, 79–84.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, *14*(1), 366–410.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, *33*(1), 126–148.

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.

Mayer, A.-S., Strich, F., & Fiedler, M. (2020). Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive*, *19*(4).

Möhlmann, M., Zalmanson, L., Henfridsson, O., & Gregory, R. W. (2021). Algorithmic management of work on online labor platfoms: When matching meets control. *MIS Quarterly*, *45*(4).

Pavone, G., Meyer-Waarden, L., & Munzel, A. (2023). Rage Against the Machine: Experimental Insights into Customers' Negative Emotional Responses, Attributions of Responsibility, and Coping Strategies in Artificial Intelligence–Based Service Failures. *Journal of Interactive Marketing*, *58*(1), 52–71.

Sniezek, J. A., & Buckley, T. (1995). Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174.

Strich, F., Mayer, A.-S., & Fiedler, M. (2021). What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees' Professional Role Identity. *Journal of the Association for Information Systems*, *22*(2), 304–324.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.*, 1–9.

Teodorescu, M., Morse, L., Awwad, Y., & Kane, G. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly*, *45*, 1483–1500.

Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of Educational Psychology*, *71*(1), 3–25.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573.