The Effect of AI Advice on Human Confidence in Decision-Making

Anna Taudien
University of Cologne
taudien@wiso.uni-koeln.de

Andreas Fügener
University of Cologne
andreas.fuegener@uni-koeln.de

Alok Gupta
University of Minnesota
gupta037@umn.edu

Wolfgang Ketter
University of Cologne
ketter@wiso.uni-koeln.de

Abstract

As artificial intelligence advances, it can increasingly be applied in collaboration with humans in decision-making contexts. However, questions on the design of different collaborative environments remain open. In the context of AI-advised human decision-making processes, we analyze the influence of AI advice on human confidence in the final decision. In a laboratory experiment, 458 subjects performed an image classification task. We compare their confidence over three treatments: i) a baseline case where subjects do not receive any AI advice; ii) where subjects receive AI advice; and iii) in addition to AI advice subjects also see the certainty of AI for its choice. Our results suggest that while AI advice can increase human overconfidence, this effect can be mitigated by augmenting the AI advice with its certainty. Our results not only contribute to the growing literature of human-AI collaboration, but also bear important practical implications for the design of collaborative systems.

1. Introduction

The importance and dispersion of artificial intelligence (AI) is increasing in many areas of everyday and professional life. AI can be summarized as systems which are able to interpret external data correctly, learn from these, and use them to achieve specific goals and tasks. In doing so, AI is able to perform perceptual, cognitive, and conversational tasks [1, 2]. Due to vast technological improvements, AI can increasingly be applied to support decision-making processes [3]. Nevertheless, due to legal [4] or ethical [5] restrictions, in many application areas it is not feasible to let AI make autonomous decisions [6]. At present, in most cases AI does not decide autonomously, but works in collaboration with humans. Empirical research has already shown that a collaboration of humans and AI can outperform both of the actors on their own [7]. Human-AI collaboration aims at combining the individual strengths of humans and AI and can create synergistic effects to realize optimized joint decision outcomes [8]. However, it is not clear how collaborative environments should be constructed to ensure effective and efficient collaboration between humans and interacting advanced computing systems such as AI [9]. This glaring research gap has been highlighted in several calls for future research to develop techniques supporting a collaboration between humans and AI [9, 10].

A growing research stream is beginning to shed light on desired characteristics of the collaborative environments by focusing on different forms of human-AI collaboration. Amongst others, these differ in the degree of decision autonomy given to the AI. Forms of collaboration include delegation from humans to AI with high autonomy for AI [11], or AI-advised decision-making, allowing for higher human autonomy [12]. We focus on the latter in this paper.

Previous research on this topic has analyzed how AI advice is influencing different performance measures such as the accuracy of a decision [6]. It was found that AI advice can also have negative consequences on the performance of collaborative decision-making when humans rely on wrong AI advice even though they would have known the correct answer. For instance, Lebovitz et al. [13] have found that AI insights can decrease the work performance of physicians as they may cause them to doubt their own diagnostic decisions and spend more time on decision-making. To gather an expansive understanding of AI-advised decision-making, it is important to not only assess whether advice is accepted and the resulting accuracy. Moreover, potential moderators on the performance need to be contemplated. By this, a holistic picture of the impact of AI-advice on decision-making can be created.

Recent research about the influence of human advice (not from AI) has acknowledged that not only the acceptance of advice, but also its effect on confidence

URI: https://hdl.handle.net/10125/79359
978-0-9981331-5-7

is of importance. By this, a broader view of the performance through advice taking is enabled [14]. In line with this, we consider human confidence in a decision made based on AI advice to be an important aspect to broaden the understanding of effective AI-advised decision-making.

Following the research literature on human confidence, a mismatch between human confidence and the actual performance is often likely. One of the most well-known findings in this context is the one of overconfidence, stating that people systematically misjudge their own performance [15].

In a different setting of human-AI collaboration, delegation from humans to AI, the issue of humans' missing awareness of the own knowledge has already been shown. Fügener et al. [11] found that their subjects were not aware of what they know and what they do not know. Hence, they could not tell when it would be wiser to let the AI solve a task, or to perform it by themselves. Thus, collaboration did not reach its full potential due to a mismatch of humans' subjective and actual knowledge.

It is conceivable that such a mismatch can also become an issue in the context of AI-advised decision-making. If AI advice distorts human confidence, the resulting impact could become harmful.

The goal of this paper is to improve the general understanding of these intriguing issues by focusing on the influence of AI advice on human confidence. We explore these issues by running an exploratory study about the influence of AI advice and certainty on human confidence. In a first step, we want to analyze the impact of pure AI advice and by this want to answer the following research question:

RQ1: *How does AI advice influence a human decision-maker's confidence about the final decision?*

One premise for a successful collaboration through AI-advised decision-making found by previous research is to have a correct mental model of the AI's error boundaries [8]. In order to analyze the effect of this awareness on human confidence, we additionally assess the impact of providing information on the AI's certainty about a suggested decision:

RQ2: *Does human confidence change when the AI's certainty is disclosed along with its decision?*

To answer these research questions, we conducted a laboratory experiment with 458 subjects who performed an image classification task. Subjects were randomly assigned to one of three treatments: The first without any AI-advice, the second with showing the AI's answer as advice, and the third in which additionally the AI's certainty is presented. Subjects had to classify the images and for each answer state their certainty about their decision.

Our findings are relevant for contexts in which humans make decisions based on AI advice and for which it is important to optimally design the interaction processes between humans and AI. They are especially relevant for situations in which relying confidently on a decision may have severe consequences if the decision is wrong. Especially, in such high-stake domains it is of importance that the human decision-maker possesses a correct degree of confidence in a decision in order to avoid time delay or even failure.

Our experiments take place in a setting with one human collaborator. However, our findings may also be relevant for group decision-making situations. Groups might work together, discuss ideas, and finally agree on a solution. In these situations, it is easy to imagine that a human with strong confidence about her choice might dominate the discussion with other humans who are not as certain. Lorenz et al. [16] found that informing humans about other humans' choice estimates increased their confidence in their own decision.

The remainder of this paper is structured as follows. In the next section, we provide an overview over research in AI-advised decision-making and human confidence. Then, we describe our experimental approach and subsequently present our findings on the effect of AI advice on human certainty. We conclude with a discussion of these results, resulting implications, and some limitations as well as suggestions for further research.

2. Related Work

2.1. AI-advised decision-making

AI-advised decision-making can be defined as settings in which a human takes recommendations from an AI partner. The human user sees the AI's recommendation and can decide to accept the suggestion, or answer differently. Thus, the AI gives a recommendation, but the human makes the final decision [17]. The areas of application range from medical diagnoses, over recidivism prediction, to assessment of creditworthiness [6].

In order to assess whether humans receiving advice from an AI perform better than humans on their own, the performance of such collaborations has been studied by previous research [8, 18]. The quality of performance can be measured as the accuracy of the decision made based on the AI's advice [19]. Fügener et al. [18] have tested how AI advice impacts human decision accuracy and unique human knowledge (what humans know, but the AI does not). They found that receiving AI advice improves the overall accuracy for individuals, but not for (larger) groups in a wisdom of crowds setting. They conclude that the loss of unique human knowledge due to AI advice harms the group performance. As an attempt to mitigate the negative effect of AI advice on unique human knowledge, they tested to additionally present the AI's certainty. However, this did not have the desired effect, but led to an overall decrease of accepting the AI's suggestion, even if the AI would have been correct. An alternative approach of providing personalized AI suggestions worked well in settings of all tested group sizes.

It can be seen that the general performance is not the only important factor when assessing a human-AI collaboration through receiving AI advice. As humans make the final decision, also subjective measures such as their preferences for advice are of relevance. Even though some AI systems provide advice with a high accuracy, humans might prefer to rely on human advice instead [19]. Whether this is the case could not entirely be clarified by empirical research so far. There exist contrasting findings concerning preferences of human compared to algorithmic advice. Results by Dietvorst et al. ([20], [21]) suggest that humans overreact to bad advice from algorithms, often called algorithm aversion. By contrast, findings by Logg et al. [22] indicate that humans rely more on identical advice from an algorithm than another human. These seemingly contradictory findings underline that it is not sufficient to analyze objective performance characteristics of AI-advised decision settings. Instead, also subjective human perceptions need to be studied as they may impact preferences to rely on AI advice and through this the collaborative performance.

One factor that has previously been stressed by academic literature is human trust in AI and its recommendation. It can be measured as the acceptance of the AI's advice as the human's final decision [23]. Knowing when to trust or distrust the AI allows the human decision-maker to apply her own knowledge and improve the outcome of a decision when the AI's recommendation is poor [8]. Regarding factors enabling successful AI-advised decision-making, valid human mental models of the AI's output reliability have been stressed [17, 6]. By providing awareness of the error boundaries of AI, they help users to know when to trust or distrust the AI's recommendation [8, 17]. This awareness enables the human decision-maker to decide when to accept or override an AI's suggestion [6]. In order to generate this awareness, Zhang et al. [8] have tested information designs that reveal information about the AI's confidence and their effect on human trust and a decision's final outcome. They found that showing the AI's confidence can improve the calibration of human trust in AI-advised decision-making, but found no significant effect on the outcome's accuracy.

Amongst others, the impact of pure AI advice as well as combined with AI's certainty has been studied on collaborative performance, human trust, and unique human knowledge. However, so far the human confidence in the decision made based on AI advice has not been studied extensively. It should be noted that trust is a related, yet distinct concept of confidence. While trust has an impact on whether a human accepts the AI's advice, confidence measures the human's certainty in the decision. The following passage provides a broad overview of the literature regarding human confidence relevant for the setting of our present research.

2.2. Human Confidence

Topics concerning human confidence as well as existing biases around it have been prominent in research of different areas and domains, including decision-making [24, 25]. We refer to confidence as the subjective probability that a decision is correct [26]. Mainly, research on confidence has focused on the relationship between subjective estimates of confidence and actual performance [15]. One possibility to validate degrees of confidence is to contemplate the calibration of confidence statements [27]. Calibration expresses whether people's degrees of confidence and the actual levels of decision accuracy are coherent [28]. As such, the calibration of a person is a measure for the quality of people's confidence [27].

If people's confidence statements are appropriate reflections of their actual performance, they are well-calibrated [29]. Empirical research has shown that usually, people's calibration about their own knowledge is poor [29]. Often, a mismatch between confidence and accuracy has been found in the empirical literature [28, 30]. Tidwell et al. [15] define confidence levels above the actual performance as overconfidence, and those below the actual performance as underconfidence.

To measure confidence experimentally, subjects are frequently asked to state their certainty about a decision [31, 32, 27]. In line with this, we use subjects' certainty as a proxy for human confidence in our experiment which is described in the following passage.

3. Methodology

3.1. Experimental setup

We use the experiment described in Fügener et al. [18], where human certainty was measured in the experiments but not discussed in the paper. In the experiment, subjects had to perform an image classification task. This rather generic task was explicitly chosen, as human subjects should be able to perform it without being trained for it, and as findings from generic tasks may also inform specialized tasks, which is not necessarily the case for contexts requiring specific training [18].
The subjects were asked to assign a focal image (e.g., of a monkey dog) to one of ten given image classes. The class name and 13 exemplary images were shown for each of the potential classes (similar to [33]). Exemplary screenshots of the experimental interface are shown in figures 1 to 3 displaying four of ten classes due to space limitations. All subjects were asked to classify the same 100 images in a randomized order that were chosen from the ImageNet database (www.image-net.org). A classification was considered as correct if the true image class was chosen (e.g., monkey dog, instead of a briard). Together with their choice of image class, subjects had to state their certainty for the given answer on a four-point scale (1/4: "Uncertain", 2/4: "Rather uncertain", 3/4: "Rather certain", 4/4: "Certain"). This measure of confidence was chosen as it was found that people find it easier to state their confidence verbally (e.g., ”I am certain”) compared to numerical measures (e.g., ”I am 80% certain”) [31].

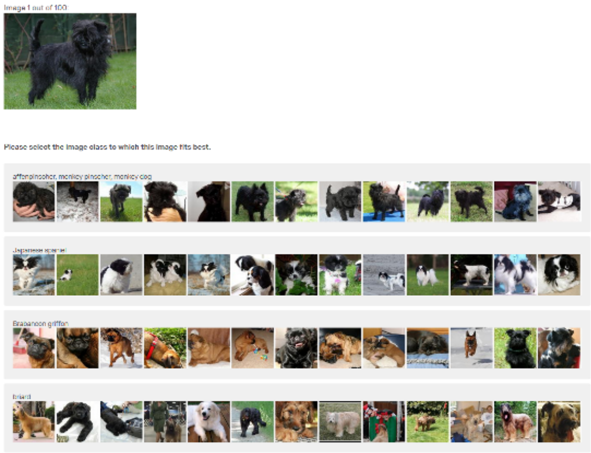
Image 1 out of 100:

Please select the image class to which this image fits best.

affenpinscher, monkey pinscher, monkey dog

Japanese spaniel

Brabancon griffon

briard

Figure 1. Exemplary screenshot of Treatment 1: Humans without AI advice.

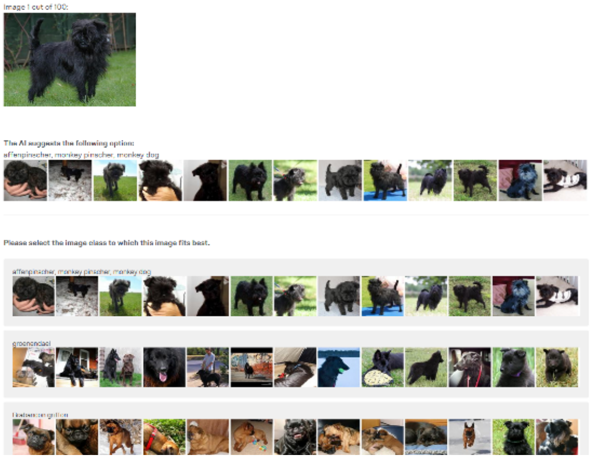
Image 1 out of 100:

The AI suggests the following option:
affenpinscher, monkey pinscher, monkey dog

Please select the image class to which this image fits best.

affenpinscher, monkey pinscher, monkey dog

groenendael

Brabancon griffon

Figure 2. Exemplary screenshot of Treatment 2: Humans with AI advice.

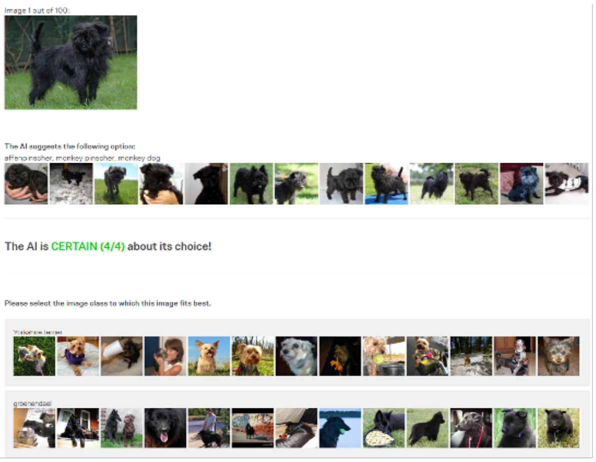
Image 1 out of 100:

The AI suggests the following option:
affenpinscher, monkey pinscher, monkey dog

The AI is CERTAIN (4/4) about its choice!

Please select the image class to which this image fits best.

Yorkshire terrier

groenendael

Figure 3. Exemplary screenshot of Treatment 3: Humans with AI advice and certainty.

Subjects were randomly assigned to one of three experimental conditions. Treatment 1 serves as a control group. Corresponding subjects had to classify all images without any help. Subjects in Treatment 2 and 3 received advice of an AI. As AI, the GoogLeNet Inception v3 one of the current best-performing AIs for image classification [34] was implemented. The AI assigns a certainty score to 1,000 potential classes representing the likelihood of a class being true for a given image. In the experiment, the AI suggested subjects the class with the highest certainty score. Two types of advice were provided: In Treatment 2, subjects were only shown the class suggested by the AI. Specifically, for a focal image they were shown which class the AI suggests above the other potential classes. Subjects in Treatment 3 were additionally shown information about the AI's certainty, which was reported on the same four-point scale used to measure human certainty. By this, subjects could easily comprehend and relate the AI's certainty to their own.

3.2. Study Protocol

The experiment was performed on August 8th, 2019, with 458 subjects recruited via Amazon Mechanical Turk ("MTurk"). Only subjects from the USA with a positive rating of at least 90%, who had not participated in related studies before were recruited. Subjects had to correctly answer an attendance check and meet technical requirements of screen resolution to be included in the study. They received $1 for participation, $0.05 for each correctly classified image, and a bonus payment of an additional $1 if they estimated how many images they classified correctly by five images. After the classification task, subjects in Treatment 2 and 3 could earn an additional $0.50 for answering a survey about their trust in AI.

Prior to the classification task, subjects were given basic information about the task and performed an attendance check. They were informed that they were not allowed to continue if they did not answer all of the questions correctly. After the attendance check, subjects were randomly assigned to one of the experimental treatments and received instructions according to it. 146 subjects were assigned to Treatment 1 ("Baseline"), 160 subjects into Treatment 2 ("Only AI suggestion"), and 152 subjects into Treatment 3 ("AI suggestion and certainty information"). After subjects classified the 100 images, they were asked to estimate how many images they classified correctly and reported how they made their decisions. Subjects of Treatments 2 and 3 additionally answered a questionnaire on human-computer trust with ten items [35]. The experiment ended with demographic questions about the subjects. Subsequently, the subjects learned about their results and payment. Average duration was 57.4 minutes and average pay without Amazon MTurk fees was $5.77.

The accuracy of the AI used by us on our chosen set of images was 77%.

4. Effect of AI Advice on Confidence

The average self-reported certainty of subjects (measured on a four-point scale: 1/4: "Uncertain", 2/4: "Rather uncertain", 3/4: "Rather certain", 4/4: "Certain") was 3.34 in the first treatment, 3.55 in the second treatment with AI suggestion, and 3.42 in the treatment where AI certainty was additionally communicated. Providing the AI suggestion to human decision makers seems to increase their level of certainty, while additionally providing the AI's certainty reduces the level of human certainty. Since the average performance in the treatments with AI advice was above the average performance without AI advice, the higher levels in certainty might be justified by a higher level of accuracy.

To explore this further, we ran a linear regression on the level of images and subjects (regression results are summarized in Table 1). The dependent variable is the level of certainty a subject reported for an image on a scale from 1 (uncertain) to 4 (certain). The independent variables are whether the AI suggestion was provided (this dummy variable was 1 for Treatments 2 and 3, and 0 for Treatment 1), whether AI certainty was provided (this dummy variable was 1 in Treatment 3, and 0 otherwise), and a dummy variable indicating whether the image was classified correctly by the human decision maker. While we just replicate the average values in Model 1, we added the interaction effect between "AI suggestion" and "Correct" as well as between "AI certainty" and "Correct" in Model 2. We control for image and subject heterogeneity with random effects.

We see a base case value of 2.961 (no AI advice, not correctly classified image). Not surprisingly, the level of subjects' certainty is significantly higher for images that were classified correctly. Providing the AI suggestion further increases confidence significantly (by 0.264) for images that were not correctly classified. This effect is mitigated by providing AI certainty which leads to a decrease of the level of certainty for incorrectly classified images by 0.347.

The interaction effects provide interesting insights. When the AI suggestions of correctly classified images are provided, the effect is negative. Thus, subjects' certainty increases less for correctly classified images as compared to incorrectly classified images. Again, providing AI certainty mitigates this effect.

We further analyze the influence of the level of AI certainty that was communicated in Treatment 3 in Model 3. In general, lower AI certainty levels led to lower human certainty levels. Thus, the communicated level of AI certainty seems to influence the decision maker's level of certainty. Interestingly, for images the human classified incorrectly, providing any AI certainty level has a negative effect on human certainty. For correctly classified images, the effect of communicating AI certainty was mostly positive, only for the lowest value of AI certainty it became negative.

Model 4 controls for the human difficulty of an image (1 minus human accuracy in T1). This replicates the observation from models 2 and 3 that providing the AI suggestion increases human certainty, especially when the classification is wrong, and that providing AI certainty mitigates the effect. We further see a negative effect of image difficulty on human certainty, and a positive interaction effect of AI suggestion and image

Table 1. Regression results on subjects' level of certainty.

	Model 1	Model 2	Model 3	Model 4
Constant	3.341***			
Constant (not correct)		2.961***	2.961***	3.482***
Correct		0.559***	0.559***	0.291***
AI Suggestion	0.206***	0.264***	0.264***	0.152**
AI Certainty	-0.131***	-0.347***		-0.247***
AI Suggestion × Correct		-0.155***	-0.155***	-0.081***
AI Certainty × Correct		0.269***		0.224***
AI Certainty				
1/4 (uncertain)			-0.394***	
2/4 (rather uncertain)			-0.341***	
3/4 (rather certain)			-0.154*	
4/4 (certain)			-0.196***	
AI Certainty (Interaction)				
1/4 × correct			-0.153***	
2/4 × correct			0.071**	
3/4 × correct			0.095	
4/4 × correct			0.298***	
Human difficulty				-1.065***
AI Suggestion × Difficulty			0.267***	
AI Certainty × Difficulty			-0.202***	
Random Effects	Images	Subject, Images	Subject, Images	Subject, Images
Adjusted R^2	0.014	0.095	0.122	0.161

Significance values: ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$

difficulty. This suggests that showing the AI suggestion offers a greater boost of confidence for difficult images. Additionally, providing the AI's certainty seems to cancel this effect. We illustrate this effect in Figure 4: the blue dots represent the effect of providing AI suggestions on human certainty for each image, while the red dots represent the effect of providing both AI suggestion and AI certainty. The regression lines illustrate that the increase in human certainty by providing AI suggestion alone increases with image difficulty (blue dotted line), while this is not the case if AI suggestion is accompanied with AI certainty (red dotted line).

Figure 2 illustrates the results. Providing an AI suggestion increases the certainty of human decision-makers in all cases, and this increase was more pronounced when humans were wrong. By providing both AI suggestion along with AI certainty, human certainty was increased if the final choice was correct, and decreased if the final choice was incorrect.

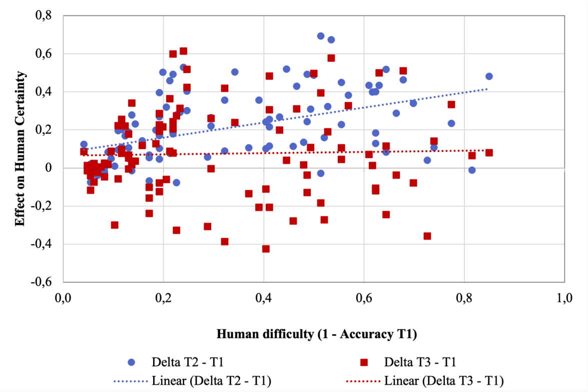

Figure 4. The effect of human image difficulty on the effect of providing AI suggestion (Delta T2-T1) and AI suggestion with AI certainty (Delta T3-T1) on the average human certainty.

5. Discussion

We have analyzed the impact of AI advice on human confidence, measured as certainty of a decision made in an image classification task. Our findings reveal that AI advice can make humans more confident about decisions, but not necessarily in a desirable way. Pure AI suggestion not only increases confidence for correctly, but also for incorrectly classified images. The magnitude of increase in human confidence is even

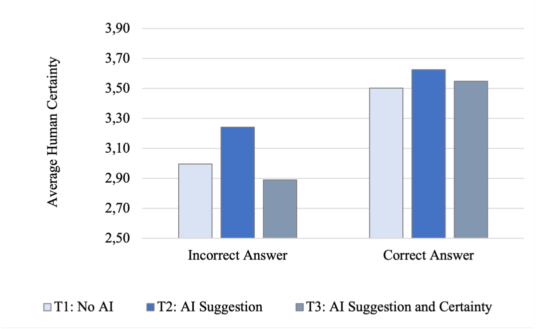

Figure 5. The effect of providing AI suggestion and certainty on average human certainty.

higher for incorrectly classified images. Thus, humans seem to feel false comfort in an incorrect decision by receiving AI advice. As such, pure AI advice tends to increase overconfidence in a decision-maker.

Additionally providing the AI's certainty for a given advice mitigates the described effect. It decreases confidence for incorrectly classified images by a much stronger magnitude than for correctly classified images. Hence, subjects achieve a better calibration regarding their confidence about a decision compared to receiving pure AI advice. Apparently, showing AI's certainty helps to reduce a part of overconfidence based on the perceived performance of AI.

This effect of mitigation might be achieved because providing the certainty of AI shows its error boundaries and helps to create a more realistic mental model of the AI [6, 8]. This can help to correct a misleading effect of seeing only pure AI advice which is also known from literature regarding advice from humans to humans. Receiving an advice from another person can (falsely) increase the own confidence in a decision, because one suspects that the other person has better information [36, 16].

5.1. Implications

The findings of this paper contribute to the growing body of literature regarding human-AI collaboration, specifically AI-advised decision-making. For the design of decision-making processes assisted by AI, our findings have important implications. In contexts where the confidence in a decision is of importance, not only pure AI advice should be presented, but also the corresponding certainty. Otherwise, increased human overconfidence could be the result. Examples for these areas are high-stake decision-making domains in which confidence in (wrong) decisions could have severe consequences [17], such as for physicians in hospitals, or brokers in the banking sector.

Our findings specifically are of special importance for decision-making scenarios in which uncertainty is involved, and where not one finite solution exists, e.g. for prediction cases. In these kinds of decision environments, certainty assessments often play a key role in arriving at a (collaborative) decision [29]. Thus, relying on a misleading feeling of confidence for arriving at a decision can be dangerous.

5.2. Limitations and Future Research

The conclusion of our findings cannot be to always provide the AI's certainty with a suggestion. Even though it seems to help with the calibration of human confidence levels for showing AI advice, it can have unwanted side-effects on other outcomes. Fügener et al. [18] conclude that the negative effect of AI advice on unique human knowledge could only slightly be mitigated by showing the AI's certainty together with an AI's suggestion. As a loss of unique human knowledge is suggested to have negative effects in group decision-making, a further analysis of the effect that providing AI advice with certainty on human confidence in group decision-making contexts should be conducted. Additionally, the effect of showing AI advice and certainty on confidence should be tested in group decision-making contexts. As confidence could affect behavior in groups, this could be valuable research area.

Our study did not take place in a specific application context to generate findings that can be generalized over different contexts of AI-advised decision-making. To test whether our findings hold true in concrete use contexts, the impact of providing AI advice and certainty on human confidence in a decision should be analyzed in practical settings. Examples may be the confidence of radiologists receiving AI advice for the interpretation of medical images [13], or of recruiters that receive AI's suggestions for which candidate to hire [37]. In both areas AI as decision support is increasingly used and making a decision relying on a false feeling of confidence is not desirable. Thus, the impact of AI advice on confidence and whether potential overconfidence can be mitigated by showing AI's certainty or even further measures would be worth studying.

In the studied setting, the AI outperforms human decision makers. The subjects could have been aware of this circumstance, what potentially drives the increase in certainty after receiving AI advice. Further research could explore whether AI advice in tasks in which humans on their own perform better than AI alone or with informational asymmetry lead to the same effects on confidence.

6. Conclusion

Our study has made a first step towards understanding the impact AI advice can have on human confidence. Resulting human overconfidence from receiving pure AI advice can get mitigated by additionally showing AI's certainty. This is of importance for decision-making contexts in which humans get assisted by AI. As both humans and AI are not perfect in decision-making in most domains, there will remain uncertainty. However, for contexts in which a human is making the final decision, calibrating human confidence is a possible lever to reduce the risk of wrong decisions based on overconfidence.

References

[1] A. Kaplan and M. Haenlein, "Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence," *Business Horizons*, vol. 62, no. 1, pp. 5–25, 2019.

[2] C. Longoni, A. Bonezzi, and C. K. Morewedge, "Resistance to medical artificial intelligence," *Journal of Consumer Research*, vol. 46, no. 4, pp. 629–650, 2019.

[3] J. M. Logg, J. A.Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90–103, 2019.

[4] J. K. C. Kingston, "Artificial intelligence and legal liability," in *Research and Development in Intelligent Systems XXXIII* (M. Bramer and M. Petridis, eds.), Proceedings of AI-2016, the Thirty-sixth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, 2016.

[5] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *nature*, vol. 563, p. 59–64, 2018.

[6] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-ai team performance," in *The Seventh AAAI Conference on HumanComputation and Crowdsourcing (HCOMP-19)*, 2019.

[7] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," 2016.

[8] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 295–305, 2020.

[9] F. Traumer, S. Oeste-Reiß, and J. M. Leimeister, "Towards a future reallocation of work between humans and machines – taxonomy of tasks and interaction types in the context of machine learning," in *International Conference on Information Systems (ICIS). Seoul, South Korea*, 2017.

[10] C. Wiethof, N. Tavanapour, and E. A. C. Bittner, "Implementing an intelligent collaborative agent as teammate in collaborative writing: toward a synergy of humans and ai," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2021.

[11] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Cognitive challenges in human-ai collaboration: Investigating the path towards productive delegation." `https://ssrn.com/abstract=3368813`, 2021.

[12] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, p. 44–56, 2019.

[13] S. Lebovitz, N. Levina, and H. Lifshitz-Assaf, "Is ai ground truth really true? the dangers of training and evaluating ai tools based on experts' know-what," *Journal of Behavioral Decision Making*, vol. 45, no. 3b, pp. 1501–1525, 2021.

[14] C. A. R. Jack B. Soll, Asa B. Palley, "The bad thing about good advice: Understanding when and how advice exacerbates overconfidence," *Management Science*, vol. 0, no. 0, pp. 1–19, Forthcoming.

[15] J. W. Tidwell, D. Buttaccio, J. S. Chrabaszcz, M. R. Dougherty, and R. P. Thomas, "Sources of bias in judgment and decision making," in *The Oxford Handbook of Metamemory* (J. Dunlosky and S. U. K. Tauber, eds.), Oxford University Press, 2016.

[16] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, "How social influence can undermine the wisdom of crowd effect," *PNAS*, vol. 108, no. 2, pp. 9020–9025, 2011.

[17] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33(01), pp. 2429–2437, 2019.

[18] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Will humans-in-the-loop become borgs? merits and pitfalls of working with ai," *MIS Quarterly*, 2021.

[19] M. Yeomans, A. Shah, S. Mullainathan, and J. Kleinberg, "Making sense of recommendations," *Journal of Behavioral Decision Making*, vol. 32, no. 4, pp. 403–414, 2019.

[20] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Consumer Research*, vol. 144, no. 1, pp. 114–126, 2015.

[21] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, vol. 64, no. 3, pp. 1155–1170, 2018.

[22] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithmic appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90–103, 2019.

[23] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *ACM Conference on Fairness, Accountability, and Transparency*, 2018.

[24] E. S. Mirko Kremer, Brent Moritz, "Demand forecasting behavior: System neglect and change detection," *Personality and Individual Differences*, vol. 57, no. 10, pp. 1827–1843, 2011.

[25] G. L. Sharp, B. L. Cutler, and S. D. Penrod, “Performance feedback improves the resolution of confidence judgments,” *Organizational Behavior and Human Decision Processes*, vol. 42, no. 3, p. 271–283, 1988.

[26] P. A., D. J., and K. A., “Confidence and certainty: distinct probabilistic quantities for different goals,” *Nature neuroscience*, vol. 19, no. 3, p. 366–374, 2016.

[27] B. Fischhoff, P. Slovic, and S. Lichtenstein, “Knowing with certainty: The appropriateness of extreme confidence,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 3, no. 4, p. 552–564, 1977.

[28] B. D. Pulford and A. M. Colman, “Overconfidence: Feedback and item difficulty effects,” *Personality and Individual Differences*, vol. 23, no. 1, pp. 125–133, 1997.

[29] S. Lichtenstein and B. Fischhoff, “Training for calibration,” *Organizational Behavior and Human Performance*, vol. 26, pp. 149–171, 1980.

[30] S. Lichtenstein, B. Fischhoff, and L. D. Phillips, “Calibration of probabilities: The state of the art to 1980,” in *Judgment under uncertainty: Heuristics and biases* (D. Kahnemann, P. Slovic, and A. Tversky, eds.), pp. 306–334, New York: Cambridge University Press, 1982.

[31] J. H. Grabman, D. K. Cash, C. R. Slane, and C. S. Dodson, “Improving the interpretation of verbal eyewitness confidence statements by distinguishing perceptions of certainty from those of accuracy,” *Journal of Experimental Psychology: Applied*, vol. Advance online publication, 2021.

[32] A. Koriat, S. Lichtenstein, and B. Fischhoff, “Reasons for confidence,” *Journal of Experimental Psychology: Human Learning and Memory*, vol. 6, no. 2, p. 107–118, 2021.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” 2015.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[35] M. Madsen and S. Gregor, “Measuring human-computer trust,” in *Proceedings of the 11 th Australasian Conference on Information Systems*, pp. 6–8, 2000.

[36] D. H. Sushil Bikhchandani and I. Welch, “A theory of fads, fashion, custom, and cultural change as informational cascades,” *Journal of Political Economy*, vol. 100, no. 5, pp. 992–1026, 1992.

[37] E. van den Broek, A. Sergeeva, and M. Huysman, “When the machine meets the expert: An ethnography of developing ai for hiring,” *MIS Quarterly*, vol. 45, no. 3b, pp. 1557–1580, 2021.