Dec 12th, 12:00 AM

Calibrating Users' Mental Models for Delegation to AI

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

Andreas Fuegener
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

Calibrating Users' Mental Models for Delegation to AI

Short Paper

Anna Taudien
University of Cologne
Albertus-Magnus-Platz,
50923 Cologne, Germany
taudien@wiso.uni-koeln.de

Andreas Fügener
University of Cologne
Albertus-Magnus-Platz,
50923 Cologne, Germany
andreas.fuegener@uni-koeln.de

Alok Gupta
Carlson School of Management
321 19th Avenue South
Minneapolis 55455, USA
gupta037@umn.edu

Wolfgang Ketter
University of Cologne
Albertus-Magnus-Platz,
50923 Cologne, Germany
ketter@wiso.uni-koeln.de

Abstract

Artificial intelligence (AI) has the potential to dramatically change the way decisions are made and organizations are managed. As of today, AI is mostly applied as a collaboration partner for humans, amongst others through delegation of tasks. However, it remains to be explored how AI should be optimally designed to enable effective human-AI collaboration through delegation. We analyze influences on human delegation behavior towards AI by studying whether increasing users' knowledge of AI's error boundaries leads to improved delegation behavior and trust in AI. Specifically, we analyze the effect of showing AI's certainty score and outcome feedback alone and in combination using a 2x2 between-subject experiment with 560 subjects. We find that providing both pieces of information can have a positive effect on collaborative performance, delegation behavior, and users' trust in AI. Our findings contribute to the design of AI for collaborative settings and motivate research on factors promoting delegation to AI.

Keywords: Artificial intelligence, delegation, human-AI collaboration, user behavior

Introduction

Delegation of tasks has long been of substantial meaning for the allocation of work in organizations. It allows for a transfer of decisions that either cannot be solved on one's own, or may free up resources to engage in other activities (Leana 1986). Due to technological advancements, delegation nowadays is no longer reserved for human-to-human interactions (Baird and Maruping 2021). Especially, the rise of artificial intelligence (AI) has facilitated delegation to information systems (IS) artifacts and comes with the promise of improved outcomes (Baird and Maruping 2021; Stout et al. 2014). Examples can be found in various domains, ranging from AI scheduling complex logistics over selecting suitable applicants for job positions to performing financial transactions on behalf of humans (von Krogh 2018).

Despite initial concerns that transferring tasks to AI may replace humans in most jobs, growing evidence suggests that having humans and AI working together may be beneficial instead (Brynjolfsson et al. 2018). One driving factor of this outcome is that humans and AI possess complementary knowledge and capabilities (Fügener et al. 2021a; von Krogh 2018). This implies that the error regions of humans and AI

in most cases do not perfectly overlap (Tan et al. 2018). If tasks between humans and AI are distributed considering these differing error regions, their complementary knowledge can be leveraged and collaborative performance should improve (Baird and Maruping 2021; Fügener et al. 2021a; von Krogh 2018). While delegation is a promising coordination mechanism, empirical studies show that humans in charge of allocating tasks between themselves and a supportive AI system fail to optimally make use of it (Fügener et al. 2021a; Zhang et al. 2020). Despite its significance, studies on the improvement of human delegation behavior to AI remain scarce (Lyytinen et al. 2021).

One approach to realize the advantage of greater knowledge through human-AI collaboration is to make the error regions of both actors visible. When humans are in charge of allocating tasks, they need to be able to assess when the AI is likely to perform well (Tan et al. 2018). In the context of humans receiving advice from AI, Bansal et al. (2019a) point out that it is vital for this assessment that humans possess a realistic mental model of the AI capabilities. The mental model in this case refers to insights about multiple aspects of the AI system's capabilities acquired by the human. The authors especially stress the importance of knowing the AI's error boundaries, namely those instances for which the AI is correct, to form a realistic mental model. A mismatch between the mental model and the AI's true error boundaries can lead to harmful decisions of either trusting the AI when it is wrong, or not trusting it when it is correct. These suboptimal decisions may lower accuracy in turn. Thereby, the success of human-AI collaboration hinges on the human's assessment when to trust the AI. Otherwise, complementarities between humans and AI will most likely not be realized (Bansal et al. 2019a; Bansal et al. 2019b).

To improve humans' mental models of AI, providing insights into the performance of an AI system has been suggested (Bansal et al. 2019b). However, simply providing feedback on the AI's performance on a task may also have a detrimental effect. Despite the high development of AI systems, they still make mistakes. Showing those mistakes to users can lead to a decrease in trust and a resulting aversion to use the AI (Dietvorst et al. 2015). One reason for this finding is that humans may have unrealistically high expectations about AI's capabilities (Glikson and Woolley 2020). In academic literature, it has been proposed to calibrate expectations of AI's performance by showing its certainty scores regarding a decision. However, there exists mixed evidence on whether this is beneficial or harmful for trust and performance in human-AI collaboration (Fügener et al. 2021b; Helldin et al. 2013; Zhang et al. 2020).

We aim to shed light on the ambiguous effects of providing feedback and confidence scores on human-AI collaboration. To our knowledge, causalities of showing either one or both pieces of information on AI's error boundaries have not been analyzed so far. Our objective is to fill this gap by exploring underlying mechanisms that may trigger beneficial or harmful effects of increased information on AI in the context of delegation. For this, we focus on trust in AI and human delegation behavior. We propose the following research question (RQ):

RQ: How does providing users with information about an AI's error boundaries influence their trust in and delegation behavior towards AI?

To answer this question, we conducted a 2x2 between-subjects online experiment with 560 participants recruited via Amazon MechanicalTurk (MTurk). We varied whether participants were shown the AI's certainty before a delegation decision to AI and whether they received feedback on the AI's outcome afterwards. Our preliminary findings show that trust, delegation behavior, and collaborative performance improve through providing both pieces of information. Further, we see that higher delegation is associated with higher collaborative performance. However, we also find that providing AI's certainty can lead to a shift of delegation behavior towards delegating easier tasks which is not optimal. We contribute to the literature on human-AI collaboration in general and on delegation to AI-based systems in particular. We provide insights on the degree of information that AI-based systems should give to a user in delegation contexts and elucidate the underlying mechanisms supporting collaborative performance. As such our findings have implications for the design of collaborative AI systems.

Theoretical Background and Related Work

Delegation in the Organizational and IS context

Delegation builds on the theory of agency analyzing how one party (called principal) transfers work to another (the agent) (Eisenhardt 1989). This transfer includes risk as it may be difficult for the principal to

first verify whether the agent is capable of solving a task and to second assess whether the agent has accomplished the task as desired (Eisenhardt 1989; Stout et al. 2014). To minimize this risk, principals have been found to be more likely to delegate when they feel agents are competent and capable (Leana 1986) as well as trustworthy (Aggarwal and Mazumdar 2008; Stout et al. 2014).

Academic studies on delegation decision-making can be found in areas including management and leadership (Leana 1986), marketing (Botti and McGill 2006), or psychology (Aggarwal and Mazumdar 2008; Botti et al. 2009). Therefore, findings and contexts of delegation from these sources are quite broad. Yet, the studies share a consistent theme of the distribution of rights and responsibilities between principals and agents as well as how coordination such as monitoring and updating occurs after an initial delegation. Such mechanisms for monitoring and updating especially include feedback loops on delegation outcomes to the principal (Baird and Maruping 2021).

Nowadays, not only humans can possess agency, but also IS artifacts such as AI-based systems which enables humans to delegate more and more complex tasks to these increasingly autonomous artifacts. Expanding agency today even allows for delegation in the opposite direction from AI to humans (Baird and Maruping 2021). Fügener et al. (2021a) experimentally analyzed delegation both from humans to AI and from AI to humans in a setting with complementary skills using an image classification task. Their findings suggest that combined performance is significantly stronger when AI distributes the work. While feedback on the AI's performance has not been central to their analyses, they included an experiment as a robustness check in which subjects received continuous feedback on their own and AI's performance. They did not find significant differences compared to humans delegating without any outcome feedback. This null result indicates that providing feedback on the AI's performance alone does not improve humans' delegation behavior to AI. In a setting without complementary skills where humans and AI show comparable performance when working on a task alone, Zhang et al. (2020) studied the effect of showing confidence scores for a particular AI prediction. While they found no effect on the combined performance in a setting of humans delegating to AI with feedback, they showed that for high confidence levels of AI (>80%), participants' trust was significantly enhanced. However, when the AI showed confidence scores below 60%, their participants seemed to trust the model less than those who did not see AI's confidence.

Trust in Automated Systems

Since delegation situations include a high degree of uncertainty and risk for the human principal, trust in AI can be decisive for the outcome of this collaboration setting (Baird and Maruping 2021; Zhang et al. 2020). Trust may be defined as the attitude that an agent will help achieve a user's goals in a situation characterized by uncertainty (Lee and See 2004). Whether a user trusts the collaborating AI system has found to be one predictor of the level of reliance on the technology. Yet, how and whether to build trust in AI is still not conclusively answered in IS research (Glikson and Woolley 2020).

To reduce issues arising from agency relationships, increasing a system's transparency has been proposed. Wang and Benbasat (2007) studied trust in the context of a web-based support technology and found that explanations aimed at decreasing information asymmetry between the principal and the agent can positively affect trust in the system. Their explanations included "how" explanations revealing a system's line of reasoning, and "why" explanations justifying the purpose of a system's actions (Wang and Benbasat 2007). However, explanations of these kind are often hard to realize for complex AI systems such that in our study we assess the provision of simpler to implement performance measures. Hence, in the following we focus on effects of increased system transparency through information on its reliability.

Recent research has stressed the importance of not contemplating trust in absolute terms. Instead, a correspondence between the extent of users' trust in an automated system and the system's capability needs to be created, a process named trust calibration (Chen et al. 2021). To enable trust calibration, providing information on the system's performance was found to be effective (Hoff and Bashir 2015). Specifically, information on a system's real-time confidence levels (Hoff and Bashir 2015) and accuracy (Chen et al. 2021) may foster trust calibration. Yet, these effects on trust calibration were shown for providing information on confidence and accuracy in isolation while we also study them in combination.

As trust is not a static, but a dynamic concept, it may change based on the agent's observed behavior. High levels of (calibrated) trust in AI-based systems may also quickly diminish once humans observe the system err (Dzindolet et al. 2003). A prominent finding named algorithm aversion describes this phenomenon in the context of human reliance on a forecasting algorithm. Despite initial high reliance,

humans tend to trust the algorithm less after they observed it making a mistake (Dietvorst et al. 2015). As such, increasing transparency of a system's performance does not always have to be beneficial.

Methodology and Results

To analyze the effects of providing more information about AI's error boundaries on users' trust and delegation behavior we conducted an online experiment. We applied a 2x2 between-subjects design with variations of showing the AI's certainty before (no certainty, certainty) and the AI's performance after (no outcome, outcome) a delegation decision. We have pre-registered our study design together with hypotheses on the effectiveness of these variations based on theory developed from academic literature (https://osf.io/mj3y9). In this paper, however, we exploratory analyze the underlying mechanisms leading to delegation behavior. Therefore, the hypotheses and their tests are not included in this work as they are out of scope for this short paper.

Experimental Task and Setting

We use an image classification task for the experiment. The general setup is similar to the experiments in Fügener et al. (2021a) but it contains different images and some adaptations. The task implies selecting an image class for a given focal image. The classes can be understood as content groups and cover a broad range of humans, animals, and objects. We chose this task for two main reasons: First, humans are naturally good at classifying images, but modern AI can perform equally well or better. Yet, they possess complementary knowledge in this task such that some images are harder for AI to classify, but comparably easy for humans and vice-versa. Second, we opted for a generalizable experimental setting. By this, we aimed to ensure that observed delegation behavior is not limited to the experimental context.

Subjects assigned 100 images taken from the ImageNet database (www.image-net.org) to one of ten given classes. For each class we displayed thirteen exemplary images and the classes' names (similar to Russakovsky et al. 2014). A classification is counted as correct if the focal image was assigned to the right class, i.e., the ground truth provided with the dataset. For each image subjects were given the option to delegate the classification decision to an AI. The subjects were informed beforehand that if they delegate, the AI's answer will be used as their own and they get paid accordingly. The AI we apply is the GoogLeNet Inception v3, a convolutional neural network trained on the ImageNet database with 1,000 classes. The Inception v3 assigns a score to the potential image classes and chooses the one with the highest score as its answer (Szegedy et al. 2015). In our image dataset it classified 78 out of 100 images correctly.

Experimental Procedure

The experimental procedure is depicted in Figure 1 and was approved by a local ethics committee. First, subjects received general instructions on the experimental task and their payment. They then had to complete a general attention check and perform an exemplary image classification to ensure they understood the task properly. Only subjects that passed the attention checks could continue with the experiment and were randomly assigned to one of the four treatments. Treatment 1 serves as our control group in which subjects were neither shown AI's certainty nor outcome feedback. Subjects assigned to treatment 2 were only shown the AI's certainty before they could decide whether to delegate an image. In treatment 3, subjects only saw feedback on the AI's outcome after their delegation decision. Subjects assigned to treatment 4 saw both the AI's certainty before and feedback on the AI's outcome after each delegation decision. An overview of the treatments is included in Figure 1. Conditional on the treatment they belonged to, subjects received further instructions on the succeeding image classification task.

All subjects were informed about the AI's overall accuracy of 78% for the used image data set. Besides this, subjects assigned to one of the *certainty* conditions (treatments 2 and 4) were explained that they would be displayed the AI's certainty score together with the option to delegate an image to the AI. They were explained that the certainty of AI ranges from 0% (uncertain) to 100% (certain) and depends on the AI's calculated likelihood that its chosen class is correct. Subjects assigned to one of the *outcome* conditions (treatments 3 and 4) received the instruction that after their decision to delegate, they would be shown whether the AI's classification was correct or not. This was irrespective of whether they decided to delegate. To ensure that subjects understood the AI's provided information, they had to answer an attention check adapted to the treatment they belonged to. Only if they gave the correct answers regarding the provided information, they could start with the main task of the experiment.

The main experimental task comprised the classification of 100 images. All subjects classified the same 100 images in a randomized order. They first classified each image on their own and stated their respective certainty. To ensure comparability between subjects' and AI's certainty, subjects indicated their certainty using a slider ranging from 0% (uncertain) to 100% (certain). After each image classification, subjects were directed to the next screen displaying the option to delegate the image classification to AI or use their own classification decision.

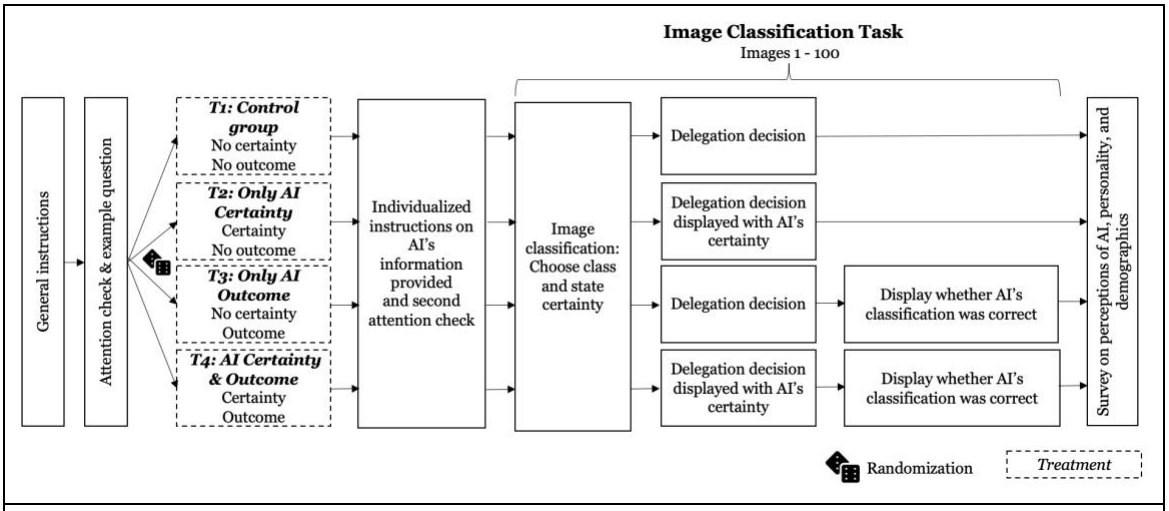

Figure 1. Experimental Flow

Subjects belonging to one of the two *certainty* conditions (treatments 2 and 4) were shown the AI's certainty on the same screen as the option to delegate. We also displayed their own stated certainty for the particular image classification which subjects entered on the preceding screen. This should ensure that they were aware of their stated certainty. If assigned to one of the *outcome* conditions (treatments 3 and 4) subjects saw another screen stating whether the classification of the AI was correct or incorrect after making their delegation decision. The information on the AI's outcome was presented with the respective focal image and the class chosen by the AI. The classification procedure was repeated for 100 images. After subjects finished the classification task, they answered a survey on their perceptions of the AI in use, personality, and demographics.

Outcome Measures

We compare the treatments based on their *collaborative performance* with AI, defined as the number of correctly classified images. For not delegated images we use the subjects' own answer, whereas for delegated images, we use the AI's answer. Further, we compare subjects' delegation behavior by analyzing the *frequency of delegation* (number of delegated images). We surveyed subjects' level of trust in AI through four items adapted from Yagoda and Gillan (2012) with a 7-point Likert rating scale ranging from "strongly disagree" to "strongly agree". To construct one measure for *trust*, we compute the mean of these items. For subjects' *calibration of trust,* we study how well they could assess whether the AI was correct for a delegation, i.e., the number of delegated images for which the AI was correct (similar to Yang et al. 2020).

Study Protocol and Descriptive Statistics

The experiment was performed on April 21 and 22, 2022. We have opted to run our study via CloudResearch which accesses the participant pool of Amazon MTurk but allowed us to add specific filters regarding the quality of recruited participants. Overall, 748 participants were recruited of which 180 were rejected by the system because they either attempted to take the survey twice (n=102), had a wrong screen resolution (n=48), or failed the attendance check (n=30). This results in 568 complete responses from which we excluded eight responses that we suspected being the same participants who took the survey in parallel. Hence, our final sample consists of 560 subjects. Subjects received a participation reward of $2.50, a performance-based pay of $0.05 for each correctly classified image, and $0.50 for filling out a

personality-related questionnaire at the end. Consequently, payment could vary between $3.00 to $8.00. Average duration of the experiment was 52:57 minutes with an average pay of $6.95.

Random assignment put 137 subjects in *treatment 1* (*no certainty, no outcome*), 144 subjects each in *treatment 2* (*certainty, no outcome*) and *3* (*no certainty, outcome*), as well as 135 subjects in *treatment 4* (*certainty, outcome*). Roughly 56% of the final sample identified as being female, almost 44% identified as being male, and one participant preferred not to report. Participants' age ranges from 18 years to the age group of 65 – 74 years with the majority indicating their age between 35 and 44 years (32%).

Results

We first provide a high-level analysis of the treatments by comparing their collaborative performance. The corresponding means in each treatment are depicted in Figure 2. In treatments 1, 2 and 4 collaborative performance exceeds that of AI alone (78%). However, the higher performance compared to AI alone is only statistically significant for the treatments in which the AI's certainty was displayed, namely treatment 2 ($t(143) = 2.05$, $p = .02$) and 4 ($t(134) = 3.72$, $p < .001$). Comparing mean differences between treatments, we only interpret values with a significance level of $p < .017$ which is the threshold after Bonferroni adjustment to correct for Type 1 errors. We find that solely the combination of providing AI's certainty and its outcome (treatment 4) leads to a significant performance improvement compared to not providing any of the information in treatment 1 ($t(277) = -2.80$, $p = .005$).

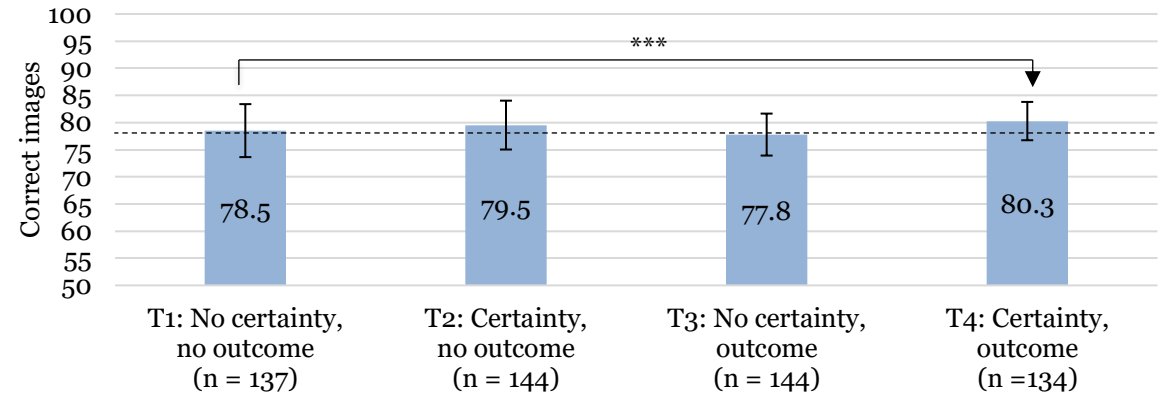

Figure 2. Average Values of Collaborative Performance, (*p < 0.01). Error bars denote standard deviations, the horizontal dashed line is the AI classification accuracy of 78%.**

In a second step, we analyze underlying reasons behind the performance exposed in the treatments. For this, we run linear regressions on the subject level with the displayed information of AI as regressors. Specifically, we analyze the influence of both showing AI's certainty and AI's outcome alone and in interaction with each other on delegation behavior, trust, and performance measures. We further assess how delegation frequency and the treatment manipulations relate to the overall performance. The results are displayed in Table 1. Subjects who did not receive any information on the AI's certainty or outcome roughly delegated 19 images. The average number of delegated images increases significantly for subjects that were only presented the AI's certainty before a delegation decision. They delegated around 5.5 images more compared to those not receiving any information. This effect is magnified by additionally providing feedback on the AI's outcome. This positive interaction effect is especially interesting because solely providing feedback decreases the frequency of delegation, but not significantly. To explore potential reasons for this effect, we take a look at humans' self-assessed trust in AI. Subjects who neither saw the AI's certainty, nor its classification outcome, showed an average trust level of 5.2 points out of 7. If only the AI's outcome was presented to participants, their average level of trust decreased significantly which may explain the directional change in delegation frequency. Yet, if displayed in combination with AI's certainty this effect is turned around.

We further split the collaborative performance, that is the number of correctly classified images, in our third and fourth model. In model 3 we contemplate how many of the delegated images were correctly classified by the AI (calibration of trust). To analyze performance among self-classified images, we assess the number of images that were not delegated and correctly classified by the human in model 4. Of the 78 images classified correctly in the control treatment, roughly 13 delegated images were correctly classified by the AI and the remaining 65 images were correctly classified by the human. When providing AI certainty, the number of delegated images correctly classified by the AI increases significantly by about 8 images, while the number of images correctly classified by the human drops by about 7. As only 5.5 images more were delegated than in the control treatment, these results indicate that providing AI certainty shifts delegation to those images which humans potentially could have correctly classified by themselves (i.e., images easy to classify for humans). While providing AI outcome feedback alone does not yield any significant changes, showing both AI outcome and AI certainty increases the number of delegated images significantly. Further, the combination of AI's certainty and outcome increases the number of images that were delegated and correctly classified by the AI. As such it positively affects the calibration of trust in AI. However, it does not significantly decrease the number of images correctly classified by the human. This explains why collaborative performance only improves when showing the information in combination. Model 5 reveals that higher delegation frequency is associated with higher overall performance (i.e., those images that were correctly classified in collaboration). While this effect is significant, the interaction between AI certainty and outcome is not, indicating that the higher collaborative performance in treatment 4 rather stems from the increased delegation frequency through showing AI certainty and outcome than directly through the treatment manipulation.

	Delegation Frequency (1)	Trust in AI (2)	Correct Images (Delegated) (3)	Correct Images (Self) (4)	Correct Images (Total) (5)
Delegation Frequency					0.044** (0.019)
AI Certainty	5.532*** (2.219)	−.0027 (0.130)	7.717*** (1.1811)	−6.685*** (1.962)	0.779 (1.011)
AI Outcome	-2.038 (2.219)	-0.266** (0.130)	-1.506 (1.811)	0.780 (1.962)	-0.643 (1.006)
AI Certainty × AI Outcome	6.232** (3.144)	0.418** (0.185)	5.109** (2.566)	-3.652 (2.780)	1.190 (1.430)
Constant	19.489*** (1.589)	5.192*** (0.093)	13.117*** (1.296)	65.387*** (1.405)	77.649*** (0.811)
Adjusted R^2	0.053	0.011	0.105	0.061	0.015
F Statistic	11.458***	3.154**	22.760***	13.162***	3.104**

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01; Standard errors are in parentheses; N = 560

Table 1. Overview of Linear Regression Models

Discussion and Conclusion

We study how providing information on AI's error boundaries affects human delegation behavior and trust in AI. While our research is still in progress, we provide first insights on underlying mechanisms of delegation behavior when humans receive information on AI's certainty before and/or feedback on AI's outcome after a delegation decision. We find partial support for the phenomenon of algorithm aversion (Dietvorst et al. 2015, 2018). Subjects who only received feedback on the AI's performance including incorrect decisions made by the AI tended to report lower levels of trust in AI than those who did not see the AI err. However, we propose that additionally providing the certainty of AI regarding a decision may serve as a mitigation of algorithm aversion. We further observe that providing the AI's certainty supports humans in relying on the AI when it is correct. As such our findings suggest that providing AI's certainty

improves the calibration of trust in AI (Yang et al. 2020; Zhang et al. 2020). However, this may not always be beneficial for performance. Our results indicate that humans who saw AI's certainty rather delegated images that were easy to classify, both for humans and AI. This explains the observed increase in performance among delegated images and simultaneous decrease of performance among self-classified images compared to subjects who did not see AI's certainty. For these images the AI has a high certainty score which may influence humans to delegate images even though they would have known the correct answer themselves. For difficult images however, the AI displays a low certainty score. This potentially has caused humans to refrain from delegating, even though those images were hard to classify for them as well. It is a fallacy to believe that images for which the AI does not show high levels of certainty should not be delegated. Instead, humans would benefit from delegating those images they are uncertain about. Regardless of the AI's performance this is a good delegation rule as delegating a task a decision-maker cannot accomplish oneself will not decrease performance (Fügener et al. 2021a).

Our findings implicate that it may not be sufficient to increase humans' understanding of the AI's error boundaries to optimize collaborative performance with AI. When a human decides about the allocation of tasks between him- or herself and the AI, the human must also be aware of which tasks can be accomplished oneself. Otherwise, the allocation of tasks cannot become optimal. This necessity has already been described by Fügener et al. (2021a) who propose that a factor hindering humans from delegating to AI is their missing knowledge about what they know and what they do not know, called metaknowledge (Evans and Foster 2011). We do find that providing more information on the AI's error boundaries improves delegation behavior, (calibration of) trust, and collaborative performance significantly. However, we propose that if humans' calibration regarding their own error boundaries would be improved, this would have an even larger effect on collaborative performance. Future research should focus on ways to improve human understanding of the own cognitive limits to improve collaboration with AI. A first step could be to support human users with assessing the difficulty of a task.

We plan to extend our research by including human confidence as another potentially influencing aspect on delegation behavior and collaborative performance with AI. We expect that this on the one hand will provide insights on potential delegation strategies used by our subjects such as delegating when the own certainty is higher than AI's certainty. On the other hand, it may enable the analysis of causalities that go beyond human perception of AI. Through path analysis, we plan to explore causalities and effect sizes of human confidence in a decision as well as trust in AI on their likelihood to delegate. This can create further insights on the importance of assessing and calibrating humans' own versus the AI's error boundaries.

Acknowledgements

We thank the Center for Social and Economic Behavior (C-SEB) of the University of Cologne for its financial support.

References

Aggarwal, P., and Mazumdar, T. 2008. "Decision Delegation: A Conceptualization and Empirical Investigation," *Psychology and Marketing* (25:1), pp. 71–93.

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

Bansal, G., Nushi, B., Kamar, E., Lasecki, W., Weld, D. S., and Horvitz, E. 2019a. "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (7:1), pp. 2–11.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. 2019b. "Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff," in *Proceedings of the AAAI Conference on Artificial Intelligence* (33), pp. 2429–2437.

Botti, S., and McGill, A. L. 2006. "When Choosing Is Not Deciding: The Effect of Perceived Responsibility on Satisfaction," *Journal of Consumer Research* (33:2), pp. 211–219.

Botti, S., Orfali, K., and Iyengar, S. S. 2009. "Tragic Choices: Autonomy and Emotional Responses to Medical Decisions," *Journal of Consumer Research* (36:3), pp. 337–352.

Brynjolfsson, E., Mitchell, T., and Rock, D. 2018. "What Can Machines Learn and What Does It Mean for

Occupations and the Economy?," *AEA Papers and Proceedings* (108), pp. 43–47.
Chen, Y., Zahedi, F. M., Abbasi, A., and Dobolyi, D. 2021. "Trust Calibration of Automated Security IT Artifacts: A Multi-Domain Study of Phishing-Website Detection Tools," Information & Management (58:1), 103394.
Dietvorst, B. J., Simmons, J. P., and Massey, C. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General* (144:1), pp. 114–126.
Dietvorst, B. J., Simmons, J. P., and Massey, C. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management Science* (64:3), pp. 1155–1170.
Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. 2003. "The Role of Trust in Automation Reliance," *International Journal of Human-Computer Studies* (58:6), pp. 697–718.
Eisenhardt, K. M. 1989. "Agency Theory: An Assessment and Review," *Academy of Management Review* (14:1), pp. 57–74.
Evans, J. A., and Foster, J. G. 2011. "Metaknowledge," *Science* (331:6018), pp. 721–725.
Fügener, A., Grahl, J., Gupta, A., and Ketter, W. 2021a. "Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation," *Information Systems Research* (Articles in Advance).
Fügener, A., Grahl, J., Gupta, A., and Ketter, W. 2021b. "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI," *MIS Quarterly* (45:3), pp. 1527–1556.
Glikson, E., and Woolley, A. W. 2020. "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals* (14:2), pp. 627–660.
Helldin, T., Falkman, G., Riveiro, M., and Davidsson, S. 2013. "Presenting System Uncertainty in Automotive UIs for Supporting Trust Calibration in Autonomous Driving," in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 210–217.
Hoff, K. A., and Bashir, M. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," Human Factors (57:3), pp. 407-434.
von Krogh, G. 2018. "Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing," *Academy of Management Discoveries* (4:4), pp. 404–409.
Leana, C. R. 1986. "Predictors and Consequences of Delegation," *Academy of Management Journal* (29:4), pp. 754–774.
Lee, J. D., and See, K. A. 2004. "Trust in Automation: Designing for Appropriate Reliance," *Human Factors* (46:1), pp. 50–80.
Lyytinen, K., Nickerson, J. V., and King, J. L. 2021. "Metahuman Systems = Humans + Machines That Learn," *Journal of Information Technology* (36:4), pp. 427–445.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. 2014. "ImageNet Large Scale Visual Recognition Challenge," *CoRR* (abs/1409.0575). (http://arxiv.org/abs/1409.0575).
Stout, N., Dennis, A. R., and Wells, T. M. 2014. "The Buck Stops There: The Impact of Perceived Accountability and Control on the Intention to Delegate to Software Agents," *AIS Transactions on Human-Computer Interaction* (6:1), pp. 1–15.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. 2015. "Rethinking the Inception Architecture for Computer Vision," *CoRR*. (http://arxiv.org/abs/1512.00567).
Tan, S., Adebayo, J., Inkpen, K., and Kamar, E. 2018. "Investigating Human + Machine Complementarity for Recidivism Predictions," *CoRR* (abs/1808.09123). (http://arxiv.org/abs/1808.09123).
Wang, W., and Benbasat, I. 2007. "Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs," *Journal of Management Information Systems* (23:4), pp. 217-246.
Yagoda, R. E., and Gillan, D. J. 2012. "You Want Me to Trust a Robot? The Development of a Human–Robot Interaction Trust Scale," *International Journal of Social Robotics* (4:3), pp. 235-248.
Yang, F., Huang, Z., Scholtz, J., and Arendt, D. L. 2020. "How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Cagliari Italy: ACM, March 17, pp. 189–201.
Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. 2020. "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*, pp. 295–305.