MIS Quarterly

SPECIAL ISSUE: MANAGING AI

WILL HUMANS-IN-THE-LOOP BECOME BORGS? MERITS AND PITFALLS OF WORKING WITH AI[1]

Andreas Fügener
Faculty of Management, Economics, and Social Sciences, University of Cologne,
Cologne, GERMANY {andreas.fuegener@uni-koeln.de}

Jörn Grahl
Faculty of Management, Economics, and Social Sciences, University of Cologne,
Cologne, GERMANY {anjoerngrahl@googlemail.com}

Alok Gupta
Carlson School of Management, University of Minnesota,
Minneapolis, MN, U.S.A {alok@umn.edu}

Wolfgang Ketter
Faculty of Management, Economics, and Social Sciences, University of Cologne, Cologne, GERMANY and
Rotterdam School of Management, Erasmus University, Rotterdam, THE NETHERLANDS {ketter@wiso.uni.koeln.de}

We analyze how advice from an AI affects complementarities between humans and AI, in particular what humans know that an AI does not know: "unique human knowledge." In a multi-method study consisting of an analytical model, experimental studies, and a simulation study, our main finding is that human choices converge toward similar responses improving individual accuracy. However, as overall individual accuracy of the group of humans improves, the individual unique human knowledge decreases. Based on this finding, we claim that humans interacting with AI behave like "Borgs," that is, cyborg creatures with strong individual performance but no human individuality. We argue that the loss of unique human knowledge may lead to several undesirable outcomes in a host of human–AI decision environments. We demonstrate this harmful impact on the "wisdom of crowds." Simulation results based on our experimental data suggest that groups of humans interacting with AI are far less effective as compared to human groups without AI assistance. We suggest mitigation techniques to create environments that can provide the best of both worlds (e.g., by personalizing AI advice). We show that such interventions perform well individually as well as in wisdom of crowds settings.

Keywords: Artificial intelligence, unique human knowledge, future of work, wisdom of crowds, analytical model, machine learning, AI–human complementarity

[1] Nicholas Berente, Bin Gu, and Radhika Santanam were the accepting senior editors for this paper. Deepa Mani served as the associate editor.

Introduction

The expectations and uncertainty about how artificial intelligence (AI) will change the workplace appear boundless. Machines now routinely do many tasks not considered amenable to automation even a decade ago. However, even if full automation becomes a technical possibility, many tasks will probably still rely on human input, as ethical (Awad et al. 2018) or legal (Kingston 2016) challenges related to fully automated systems remain unresolved. Also, because humans routinely provide complementary capabilities to algorithms, combining humans with machines potentially leads to superior outcomes.

We contribute to the growing field of AI-advised human decision making where humans receive an AI-based suggestion before making final decisions. Possible applications are discussed in Bansal et al. (2019a) and include medical decision making (e.g., Bayati et al. 2014 on the prediction of readmission rates to support physicians) or legal support (e.g., Angwin et al. 2016 on recidivism prediction to support judges). The focus of this stream of literature is to maximize performance. In contrast, we focus on an effect that has been neglected so far, that is, the implications of AI advice on unique human knowledge (i.e., the knowledge a human has, but the AI does not). Unique human knowledge has a positive effect on many collaborative work environments (Krishnan et al. 1997; Nijstad and Stroebe 2006; Paulus and Brown 2007). As a proof of concept, we test the consequences of the effect of AI advice on unique human knowledge by considering crowd-based aggregation mechanisms, which Surowiecki (2004) referred to as the "wisdom of crowds." Examples of successful wisdom of crowds applications include financial forecasting (Kelley and Tetlock 2013) and prediction markets (Wolfers and Zitzewitz 2004).

Unique human knowledge results in complementarity between humans and AI. Its existence is widely accepted in the literature on AI-advised human decision making (e.g., Bansal et al. 2019a, Bichler et al. 2010, Tan et al. 2018, and Zhang et al. 2020). The possibility that humans working with AI lose their unique knowledge (i.e., their complementarity) can be extremely detrimental to long term performance, continuous improvement, and innovation. According to Paulus et al. (2019), complementarity is essential for groups to arrive at innovative and productive ideas. Exposure to diverse perspectives can increase creativity (Nijstad and Stroebe 2006; Paulus and Brown 2007), and even top-management performance correlates with complementarity (Krishnan et al. 1997). Page (2007) models the value of diversity for collective knowledge in the so-called diversity prediction theorem, where collective performance is stated as a function of individual performance and prediction diversity. However, when humans lose this complementarity by losing their unique knowledge, the decisions of each individual human start mirroring those of other humans *and* that of the AI. Thereby, humans start acting more like machines or cyborgs ("Borgs"); they strive for perfection but are only as perfect as the AI algorithm with which they are working. Humans thus lose their ability to improve upon the mechanistic decision making, based on past observations and the data that AI decisions are based on. We demonstrate that one impact of this loss of uniqueness makes humans less effective as a group in the wisdom of crowds environments.

Modern AI systems are based on training data observed from practice and are not explicitly based on human-defined rules. Thus, AI decision rules that are derived from data differ from human decision rules, leading to structural complementarities between humans and AI. Therefore, there are task instances where a human performs better than AI due to "unique human knowledge." However, losing this unique human knowledge makes humans risk becoming Borgs (i.e., working to produce similar outcomes for similar problems without bringing their unique perspective to bear).

We explore the nature of this loss and its effect by examining the following three broad research questions both analytically and experimentally:

- How does AI advice impact human decision accuracy and unique human knowledge?
- How can the negative effect of AI advice on unique human knowledge be mitigated?
- What are the consequences of a loss of unique human knowledge on the wisdom of crowds?

We develop an analytical modeling framework to develop and support our hypotheses from the perspective of rational decision makers. We then perform a series of experiments to demonstrate that our theoretical insights are empirically supported. While we recognize that, post factum, a host of theoretical frameworks can support similar results, it is quite remarkable that a simple analytical framework can highlight such a critical finding that has been largely ignored in the literature.

Our key insight is that AI advice can decrease the complementary knowledge between humans and the AI, particularly the unique knowledge of humans, even if the system successfully lifts individual human performance above the

level of a superior AI. This result enables a novel and unexplored understanding of the effects of AI advice on decision making.

We test two different interventions to mitigate the decrease in unique human knowledge:

- Presenting the AI's certainty of its suggestion to human decision makers. Since a modern AI can determine its own uncertainty (Zhang et al. 2020) and quantify its accuracy (e.g., confidence interval, probability of correct identification), we can communicate AI certainty to humans. Such an intervention should help humans react better to AI advice by providing a better differentiation between correct and incorrect AI advice. Presenting AI certainty improves human perception of the AI's error boundary (Bansal et al. 2019b, p. 2, refer to it as considering "when does the AI err?"). Therefore, providing AI certainty should mitigate the reduction in unique human knowledge since humans are then in a better position to ignore incorrect AI suggestions.

- Provide personalized AI suggestions to human decision makers. Again, coupling the AI's ability to judge its own certainty with the modern AI's ability to learn from human behavior, this intervention personalizes the help to human decision making by monitoring individual reactions to correct and incorrect advice and then selectively providing the advice. In other words, the AI estimates the benefit of correct advice and negative impact of incorrect advice for a particular individual. It then determines whether to make suggestions.

Our results indicate that both interventions managed to mitigate the decrease in unique human knowledge.

We then investigate the impact of AI advice on collective decision making by investigating a wisdom of crowds setting. We demonstrate a serious pitfall of AI advice: while the AI advice improves the performance of individuals and smaller groups, it significantly harms the performance of larger groups. When we test the impact of our interventions that mitigate the decrease in unique human knowledge, our first intervention does not fully overcome the pitfalls of AI advice to larger groups. While the decrease of unique human knowledge is mitigated by communicating AI certainty, the performance of crowds in which all individuals received AI certainty was still inferior to the performance of crowds in which no individual received any AI suggestions. On the other hand, our second intervention seems to be quite promising. Humans who received personalized AI suggestions performed well both individually as well as in groups of all sizes.

The remainder of this paper is structured as follows. First, we review the most current advances in the field of AI-advised human decision making. We also summarize the foundations of wisdom of crowds mechanisms and the antecedents of performance of such approaches. The subsequent section presents an analytical model explaining the effects of AI advice on human accuracy and unique human knowledge, the effectiveness of our two interventions, and possible results of applying these interventions in wisdom of crowds settings. We then present the details of our three experimental studies and the individual-level results. Following the individual level results, we present the consequences of AI advice and the effectiveness of our intervenetions in wisdom of crowds settings. Finally, we conclude with a summary of results, their implications, and the limitations of this study.

Literature Review

In this section, we discuss recent works on AI-advised human decision making, especially with respect to integrating measures of AI certainty. We then summarize the main attributes of wisdom of crowds mechanisms. The effect of advanced information technologies on human decision making can be traced back to the seminal framework presented by Huber (1990) and empirically discussed by Leidner and Elam (1995). Huber generally states that computer-assisted decision-support enables better decisions based on two dimensions: on an organizational level, the use of advanced information technologies enables a higher level of organizational intelligence, while on a decision-making level, the quality and quantity of information sources and the focus on decision making itself is positively affected. While this past research identifies complementary skills of technology, such as the ability to better apply organizational processes (without actually labeling it as such), we are, however, not aware of any research that explicitly discusses the effect of advice on performance-enhancing complementarities as they pertain to the execution of tasks.

Recently, a literature stream has evolved around AI-advised human decision making, where "a user takes action recommendations from an AI partner for solving a complex task" (Bansal et al. 2019a, pp. 2429). A vibrant topic in this research stream is whether including additional information about AI advice, such as certainty measures, increases performance. To ensure humans can assess AI's uncertainty, Zhou and Chen (2019) propose a stylized framework for integrating uncertainty in human–AI decision-making

environments by creating an uncertainty-performance-interface. In a similar vein, Bansal et al. (2019a) and Bansal et al. (2019b) state that humans' mental model of the AI's error boundary is crucial for realizing complementarities between humans and AI. They designed a set of abstract experiments to test whether humans can estimate the error bounds of a system and conclude that AI systems need to be compatible with humans' mental model. In other words, a system should consistently perform well in conditions in which human expectation is that the system will perform well. If a system is created or updated in an incompatible way—one in which it does not perform well against human expectation—the overall human–AI system performance suffers. Studies that test the results of AI providing its certainty (or related measures) lead to inconclusive results. In a deception detection experiment, where genuine and deceptive hotel reviews have to be classified, Lai and Tan (2019) find that adding explanatory information to AI suggestions significantly increases performance. However, even in the best case, the AI-advised human accuracy remains below pure AI performance. In a follow-up study, Lai et al. (2020) show a use case in which training and tutorials for users lead to an improved performance of humans. Nonetheless, in their study, AI-advised human accuracy remains below pure AI accuracy by a large margin as well. Thus, the beneficial effects of additional information related to AI certainty on AI-advised human accuracy could be explained by an increased adherence with AI advice. In a case study in which AI confidence scores and explanations were added to AI suggestions in a prediction task, Zhang et al. (2020) do find indications that humans rely more on the AI when confidence is communicated, but do not see a significant increase in final prediction accuracy. A similar "null" result was found by Carton et al. (2020), who leverage interpretable machine learning algorithms to explain predictions of toxic online behavior detection tasks to human decision makers. Presenting AI suggestions biases humans toward the AI prediction; however, adding an explanation does not lead to significant effects on user accuracy or agreement to AI predictions. We conclude that the literature consistently shows a benefit of AI advice on accuracy. While there are theoretical indications that providing AI certainty to human decision makers should increase accuracy, there is mixed evidence in experimental studies. However, to the best of our knowledge, there is no study discussing the effect of AI advice on unique human knowledge. Unique human knowledge is particularly important in decision environments including groups of humans, such as wisdom of crowds.

The wisdom of crowds (Surowiecki 2004) describes the collective opinion or knowledge of a group of people achieved by aggregation of individual knowledge. A popular example is Galton's report of an ox weight-judging competition with approximately 800 participants (Galton 1907). The *vox populi*, in this case the median issued weight of the ox, was within 1% of the truth. Many follow-up studies demonstrate the power of simple crowd aggregation algorithms considering means, medians or modal choices. In a study on the value of diversity in simulations of artificial problem solvers, Hong and Page (2004) demonstrate that combining a variety of different agents might outperform combining just the best-performing, yet similar, agents. Page (2008) summarizes these findings, connects them to the wisdom of crowds, and develops the diversity prediction theorem. It states that the error of a group is positively correlated with the individual error and negatively correlated with the diversity of its members. Hong et al. (2016) synthesized three necessary conditions to realize the full potential of crowds: members should (1) hold diverse opinions (2) make independent decisions, and (3) have their own local and decentralized knowledge sources. Only few studies mention situations in which a reduction of diversity due to a concentration of expertise goes hand in hand with an increase in crowd performance. By putting higher weights on individuals with high relative performance and lower weights on individuals with low relative performance, Budescu and Chen (2015) achieve superior results. A main factor that decreases diversity is communication between crowd members. Lorenz et al. (2011) show negative effects of social influence on diversity and on crowd performance. Crowds in which subjects can reconsider initial estimates after being informed of other responses perform worse. A more nuanced conclusion stems from Becker et al. (2017), where decentralized communication networks among crowd members could increase crowd performance, whereas centralized networks led to detrimental effects. By eliminating public knowledge in crowd stock forecasting, Da and Huang (2020) could increase diversity among subjects and resulting crowd performance. We conclude that most research on wisdom of crowds validates the diversity prediction theorem: decreasing diversity, that is, a decreasing level of unique knowledge, typically leads to a decrease in crowd performance. However, we are not aware of studies that test the effect of individual (AI-based) advice on crowd diversity, individual performance, and crowd performance. Next, we develop a theoretical analytical framework to generate predictions and testable hypotheses for the AI-advised decision making setting.

Theoretical Framework: Individual Decision Making

We propose a theoretical framework based on a simple discrete choice model. Please find a summary of the relevant

notation in the Appendix. We first set up a model for the computation of baseline performance of human accuracy, AI accuracy, unique human knowledge, and unique AI knowledge. We then consider the impact of our interventions on these factors. Based on the theoretical results, we derive testable hypotheses. For individual effects, we have two sets of hypotheses: (1) impact on human performance and (2) impact on unique human knowledge. As we discuss different treatments, we derive hypotheses regarding the impact of the interventions on both measures.

Let us now provide our model description. A task $t \in \mathcal{T} = \{1..T\}$ is to select the correct choice out of $c \in \mathcal{C} = \{1..C\}$ possible choices, where, without loss of generality, $c \in \{1\}$ represents the correct choice, and $c \in \{2..C\}$ represent incorrect choices. We further define p_{tc} to be the probability that a human selects choice c in task t, where $\sum_{c \in \mathcal{C}} p_{tc} = 1$. Therefore, a human selects the correct choice for a task t with probability p_{t1}.

The expected performance of a human over a set of tasks $t \in \mathcal{T}$ is

$$\frac{1}{T}\sum_{t \in \mathcal{T}} p_{t1} \tag{1}$$

Next, we characterize the nature of AI advice. For each possible choice c of a task t, the AI estimates a likelihood l_{tc}^{AI} of being correct. AI chooses the option with the highest likelihood and communicates its advice to a human. We define the AI advice a_{tc} to be 1, if the AI recommends choice c at task t, and 0 otherwise. Thus, the performance of the AI over a set of tasks $t \in \mathcal{T}$ can be defined as:

$$\frac{1}{T}\sum_{t \in \mathcal{T}} a_{t1} \tag{2}$$

Note that the definition of human and AI performance differs: While human choice is probabilistic, that is, they select a certain choice with a certain probability, the AI always selects the choice with the highest likelihood l_{tc}^{AI}. Thus, the expected human performance equals the average probability of selecting the correct choice over all task p_{t1}, while the AI performance equals the sum of selected correct choices divided by the number of tasks. We now analytically characterize the complementary knowledge between a human and the AI. We define "unique AI knowledge" (UAK) for the case where the AI is correct ($t \in \mathcal{T}^{AI}$) and a human is incorrect. Similarly, we define "unique human knowledge" (UHK) for the case where the AI is incorrect ($t \in \mathcal{T}^{\overline{AI}}$) and a human is correct. Thus, for a given human, over a set of tasks $t \in \mathcal{T}$, we can define expected unique AI knowledge (UAK) and expected unique human knowledge (UHK):

$$\begin{aligned} UAK &= \frac{1}{T}\sum_{t \in \mathcal{T}} \max(0, a_{t1} - p_{t1}) \\ &= \frac{1}{T}\sum_{t \in \mathcal{T}^{AI}} 1 - p_{t1} \end{aligned} \tag{3}$$

$$UHK = \frac{1}{T}\sum_{t \in \mathcal{T}} \max(0, p_{t1} - a_{t1}) = \frac{1}{T}\sum_{t \in \mathcal{T}^{\overline{AI}}} p_{t1} \tag{4}$$

Note that $UAK - UHK = \frac{1}{T}\sum_{t \in \mathcal{T}^{AI}}(1 - p_{t1}) - \frac{1}{T}\sum_{t \in \mathcal{T}^{\overline{AI}}} p_{t1}$ $= \frac{1}{T}\sum_{t \in \mathcal{T}} a_{t1} - \frac{1}{T}\sum_{t \in \mathcal{T}} p_{t1}$, which is the difference between AI and human performance. The expected common knowledge (tasks that both a human and the AI solve correctly) is $\frac{1}{T}\sum_{t \in \mathcal{T}^{AI}} p_{t1}$, and the expected missing knowledge (tasks that neither a human nor the AI solve correctly) is $\frac{1}{T}\sum_{t \in \mathcal{T}^{\overline{AI}}}(1 - p_{t1})$. We denote $\frac{UHK}{UAK}$ as relative unique human knowledge.

In the following subsections, we focus on two measures of interest: First, we consider effects on final performance, and second, we consider effects on unique human knowledge. We focus specifically on the latter as unique human knowledge is an indicator for the value a human can contribute to collaborative work environments.

Effect of AI Advice on Performance

We assume that the AI advice increases the probability that a human will select the recommended choice, and decreases the probability of a human selecting any other choice. We denote the effect of AI advice on a human decision for task t as $e_t \in (0, \infty)$. The probability of selecting choice c after receiving AI advice p_{tc}^{AI} can be defined as follows:

$$p_{tc}^{AI} = \frac{p_{tc} + a_{tc} \cdot e_t}{1 + e_t} \tag{5}$$

If $e_t = 0$, the AI advice is ignored. The larger the value of e_t, the more the advice is followed by a human. The effect of AI advice on the performance of an individual for task t can be derived as

$$\Delta AI_t = (a_{t1} - p_{t1})\frac{e_t}{1 + e_t} \tag{6}$$

which is positive for $a_{t1} = 1$ and negative for $a_{t1} = 0$. We refer to this as benefit of correct advice ($a_{t1} = 1$) and harm of incorrect advice ($a_{t1} = 0$).

The total effect of AI advice on expected human performance over a set of tasks $t \in \mathcal{T}$ can be derived as

$$\Delta AI = \frac{1}{T}\sum_{t\in\mathcal{T}}(a_{t1} - p_{t1})\frac{e_t}{1+e_t} \quad (7)$$

Let e^{AI} represent the strength of effect of AI advice for all tasks where the AI advice is correct, that is, $e_t = e^{AI}$ for $t \in \mathcal{T}^{AI}$, and $e^{\overline{AI}}$ represent the strength of effect of AI advice for all tasks where the AI advice is incorrect, that is, $e_t = e^{\overline{AI}}$ for $t \in \mathcal{T}^{\overline{AI}}$. Let δ_e denote the scaling factor on the effect of correct advice relative to incorrect advice such that $\delta_e \frac{e^{\overline{AI}}}{1+e^{\overline{AI}}} = \frac{e^{AI}}{1+e^{AI}}$. Then, we can define the total effect of AI advice on human performance a set of tasks $t \in \mathcal{T}$ in terms of UAK and UHK as follows:

$$\Delta AI = \delta_e \frac{e^{\overline{AI}}}{1+e^{\overline{AI}}}UAK - \frac{e^{\overline{AI}}}{1+e^{\overline{AI}}}UHK \quad (8)$$

$$\delta_e > \frac{UHK}{UAK} \Rightarrow \Delta AI > 0 \quad (9)$$

This leads to our first theoretical proposition.

Proposition 1. The overall performance effect of AI advice is positive, if the scaling factor of the effort of correct advice δ_e is greater than the relative unique human knowledge $\frac{UHK}{UAK}$.

Thus, the effect of AI advice depends on the potential of (correct) AI advice (UAK), the risk of (incorrect) AI advice (UHK), and the human ability to differentiate between correct and incorrect advice (δ_e). Depending on the effect strength of correct and incorrect advice, the total effect of AI advice on human performance lies between minus UHK, that is, losing all unique human knowledge, and UAK, that is, gaining all unique AI knowledge. If δ_e is greater than one (that is, if correct advice has a stronger effect than incorrect advice), AI advice can be beneficial even if AI performance is worse than human performance.

Our first hypothesis makes the following assumptions that are commonly made in the literature:

i) Complementarities exist ($UHK, UAK > 0$) (see Tan et al. 2018 or Zhang et al. 2020);

ii) AI accuracy is at least on par with human accuracy ($UAK \geq UHK$) (see Russakovsky et al. 2015 for image classification challenges); and

iii) Humans are able to differentiate between correct and incorrect advice ($\delta_e > 1$) (see Bonaccio and Dalal 2006 for human advice taking, and Bansal 2019b for AI advice).

Note that if (i) and (iii) apply, it is not necessary that AI accuracy exceeds human accuracy. Still, AI algorithms outperform humans in a growing number of applications, including the automatic detection of skin cancer (Esteva et al. 2017) or games like Poker (Moravčík et al. 2017) and Go (Silver et al. 2016).

Hypothesis 1a: Human accuracy increases when receiving AI advice.

Effect of AI Advice on Unique Human Knowledge

We denote the unique human knowledge after receiving AI advice as UHK^{AI}, and the effect of AI advice on unique human knowledge as ΔUHK^{AI}. Based on our analytical framework, these can be derived as

$$UHK^{AI} = \frac{1}{T}\sum_{t\in\mathcal{T}^{\overline{AI}}} p_{t1}^{AI} \quad (10)$$

$$\Delta UHK^{AI} = -\frac{1}{T}\sum_{t\in\mathcal{T}^{\overline{AI}}} p_{t1}\frac{e_t}{1+e_t} \quad (11)$$

To provide the intuition for our second hypothesis, we present a simplification in the expression for ΔUHK^{AI} using $e^{\overline{AI}}$:

$$\Delta UHK^{AI} = -\frac{e^{\overline{AI}}}{1+e^{\overline{AI}}}UHK \quad (12)$$

Thus, AI advice decreases UHK for any positive effect of incorrect AI advice. In other words, unless humans have a perfect ability to distinguish between correct and incurrect AI advice, the AI advice reduces unique human knowledge. Therefore, we can derive our second testable hypothesis:

Hypothesis 1b: Unique human knowledge decreases when receiving AI advice.

Both the merits of AI advice on human performance and the pitfalls of AI advice on unique human knowledge relate to the effect of AI advice on human decision making, and, especially, human capability in distinguishing between correct and incorrect advice. Next, we derive the conditions

for two possible interventions that may reduce the decline in unique human knowledge:

1) The AI provides its certainty along with the advice. This can be seen as the case where humans react to the AI based on its certainty. As discussed earlier, knowing AI certainty may allow humans to better distinguish between correct and incorrect advice.

2) The AI learns about the effect of its advice on a human's decision making, and decides whether an individual should be provided a suggestion or not for each task. This is essentially equivalent to the case where the AI reacts to humans based on its assessment of human capability.

Providing AI's Certainty

To model the impact of providing the AI's level of certainty (AI certainty) to humans, we introduce a scaling parameter $s_t \in (0, \infty)$ that describes the change in the effect of AI advice, e_t, if AI certainty is communicated, leading to the following probability of selecting choice c at task t:

$$p_{tc}^{AI-cert} = \frac{p_{tc} + a_{tc} \cdot s_t \cdot e_t}{1 + s_t \cdot e_t} \tag{13}$$

Consequently, the effect of receiving AI certainty on expected performance of task t is

$$\Delta cert_t = (a_{t1} - p_{t1}) \frac{e_t}{(1 + e_t)} \frac{s_t - 1}{(1 + s_t \cdot e_t)} \tag{14}$$

Note that $\Delta cert_t$ measures the additional effect of providing AI certainty on top of providing AI advice. The expression above reflects that communicating AI certainty for task t has a positive effect on the expected performance in the following cases:

1) $a_{t1} - p_{t1} > 0$, that is, the AI advice is correct, and $s_t - 1 > 0$, that is, receiving AI certainty increases the effect.

2) $a_{t1} - p_{t1} < 0$, that is, the AI advice is incorrect, and $s_t - 1 < 0$, that is, receiving AI certainty decreases the effect.

In all other cases communicating AI certainty has either no effect or a negative effect on human performance.

Effect on Human Performance

The total effect of receiving AI certainty on human performance over a set of tasks $t \in \mathcal{T}$ can then be written as

$$\begin{aligned} &\Delta cert \\ &= \frac{1}{T} \sum_{t \in \mathcal{T}^{AI}} (1 - p_{t1}) \frac{e_t}{(1 + e_t)} \frac{s_t - 1}{(1 + s_t \cdot e_t)} \\ &- \frac{1}{T} \sum_{t \in \mathcal{T}^{\overline{AI}}} p_{t1} \frac{e_t}{(1 + e_t)} \frac{s_t - 1}{(1 + s_t \cdot e_t)} \end{aligned} \tag{15}$$

Again, we consider an effect strength of advice of e^{AI} for tasks where the AI advice is correct, and of $e^{\overline{AI}}$ for tasks where the AI advice is incorrect. We further consider a scaling effect of communicating AI certainty of s^{AI} for tasks where the AI advice is correct, and of $s^{\overline{AI}}$ for tasks where the AI advice is incorrect. Analogous to the relative effect strength δ_e, we denote δ_s as a scaling factor on the effect of receiving the AI's certainty for correct advice relative to incorrect advice with $\delta_s \frac{s^{\overline{AI}} - 1}{1 + s^{\overline{AI}} \cdot e^{\overline{AI}}} = \frac{s^{AI} - 1}{1 + s^{AI} \cdot e^{AI}}$. While δ_e symbolizes the human ability to differentiate between correct and incorrect advice, δ_s symbolizes the human ability to react differently to low and high AI certainty values (assuming that correct advice is associated with higher certainty values). Note that a negative value of $s^{\overline{AI}} - 1$ relates to a reduction of the effect of AI advice.

$$\begin{aligned} \Delta cert = \frac{e^{\overline{AI}}}{1 + e^{\overline{AI}}} \frac{s^{\overline{AI}} - 1}{1 + s^{\overline{AI}} \cdot e^{\overline{AI}}} (\delta_e \delta_s \, UAK \\ - UHK) \end{aligned} \tag{16}$$

i) $$\delta_e > \frac{1}{\delta_s} \frac{UHK}{UAK} \Rightarrow \Delta cert > 0 \text{ if } s^{\overline{AI}} - 1 < 0, s^{AI} - 1 > 0, \delta_s < 0 \tag{17}$$

ii) $$\delta_e < \frac{1}{\delta_s} \frac{UHK}{UAK} \Rightarrow \Delta cert > 0 \text{ if } s^{\overline{AI}} - 1 < 0, s^{AI} - 1 < 0, \delta_s > 0 \tag{18}$$

iii) $$\delta_e > \frac{1}{\delta_s} \frac{UHK}{UAK} \Rightarrow \Delta cert > 0 \text{ if } s^{\overline{AI}} - 1 > 0, s^{AI} - 1 > 0, \delta_s > 0 \tag{19}$$

We can now derive our second theoretical proposition.

Proposition 2. The overall performance effect of providing AI certainty is positive in the following three cases:

i) Providing AI certainty decreases the effect of incorrect AI advice, and increases the effect of correct advice ($\delta_s < 0$): Then, for a positive performance effect, unique human knowledge has to be greater than zero.

ii) Providing AI certainty decreases the effect of incorrect AI advice, and decreases the effect of correct advice ($\delta_s > 0$): Then, for a positive performance effect, the scaling factor of the effort of correct advice (δ_e) has to be smaller than the relative unique human knowledge weighted by one over the effect of communicating certainty (δ_s).

iii) Providing AI certainty increases the effect of incorrect AI advice, and increases the effect of correct advice ($\delta_s > 0$): Then, for a positive performance effect, the scaling factor of the effort of correct advice (δ_e) has to be greater than the relative unique human knowledge weighted by one over the effect of communicating certainty (δ_s).

Obviously, providing the AI's certainty increases performance if it increases the effect of correct advice and decreases the effect of incorrect advice as in case (i). If providing the AI's certainty decreases both the effect of correct and incorrect AI advice as in case (ii), the total effect benefits from low levels of human ability to differentiate between correct and incorrect advice δ_e and high levels of relative unique human knowledge. If providing the AI's certainty increases both the effect of correct and incorrect AI advice as in case (iii), the total effect benefits from high levels of human ability to differentiate between correct and incorrect advice δ_e and low levels of relative unique human knowledge. Note that high levels of relative unique human knowledge indicate high risk and small potential of AI advice, while low levels of relative unique human knowledge indicate low risk and large potential of AI advice.

According to Bansal et al. (2019b), reporting AI certainty enables humans to better assess the AI's error boundaries, that is, to better distinguish between cases where the AI is correct or incorrect. This corresponds to condition (i) above, that is, reporting AI certainty should decrease the effect of AI advice when it is incorrect, and increase the effect of AI advice when it is correct. We can now base our next hypothesis related to the impact of AI advice on human accuracy when receiving AI certainty for its choice.

Hypothesis 2a: Human accuracy increases when receiving the AI's certainty.

Effect on Unique Human Knowledge

We denote the total effect of receiving the AI's certainty on unique human knowledge over a set of tasks $t \in \mathcal{T}$ as ΔUHK^{cert}:

$$\Delta UHK^{cert} = \frac{1}{T} \sum_{t \in \mathcal{T}^{\overline{AI}}} p_{t1} \frac{e_t}{(1+e_t)} \frac{1-s_t}{(1+s_t \cdot e_t)} \tag{20}$$

We consider the effect of incorrect AI advice $e^{\overline{AI}}$, and the effect of incorrect AI advice $s^{\overline{AI}}$:

$$\Delta UHK^{cert} = \frac{e^{\overline{AI}}}{1+e^{\overline{AI}}} \frac{1-s^{\overline{AI}}}{1+s^{\overline{AI}} \cdot e^{\overline{AI}}} UHK \tag{21}$$

As $\frac{e^{\overline{AI}}}{1+e^{\overline{AI}}} > 0$ and $1 + s^{\overline{AI}} \cdot e^{\overline{AI}} > 0$, the effect on unique human knowledge is positive if $s^{\overline{AI}} < 1$, that is, if communicating AI certainty decreases the effect of AI advice for incorrect advice. With the same reasoning as above, we derive the next hypothesis.

Hypothesis 2b: Unique human knowledge increases by providing the AI's certainty.

Personalized Suggestions

The second intervention assumes that the effect of AI advice differs among individual humans $h \in \mathcal{H} = (1 \dots H)$. Humans might differently benefit from correct advice or be harmed by incorrect advice. The AI decides based on the individual human h and the task t whether to provide advice. We denote $d_{th} \in \{0,1\}$ as this decision. We assume that personalized AI suggestions do not differ from regular AI suggestions when they are made. Thus, the probability of selecting choice c is

$$p_{tch}^{AI-per} = \frac{p_{tc} + d_{th} \cdot a_{tc} \cdot e_{th}}{1 + d_{th} \cdot e_{th}} \tag{22}$$

Consequently, the effect of providing personalized advice (versus always providing the advice) on expected performance of task t is

$$\Delta per_{th} = (a_{t1} - p_{t1h}) \frac{e_{th}}{(1+e_{th})} \frac{d_{th}-1}{(1+d_{th} \cdot e_{th})} \tag{23}$$

Note that Δper_{th} measures the additional effect compared to providing regular AI advice as in (5). The effect of personalized advice concentrates on tasks where $d_{th} = 0$, that is, where no advice is provided. The effect is beneficial in cases of incorrect advice, and detrimental in cases of correct advice. Note that the AI likelihood l_{tc}^{AI} estimates the probability that the advice is correct. The AI has to decide whether to withhold the advice if the expected benefit of providing correct advice $l_{tc}^{AI}(1 - p_{t1h})\frac{e_{th}^{AI}}{1+e_{th}^{AI}}$ does not exceed the expected harm of providing incorrect advice $(1 - l_{tc}^{AI})p_{t1h}\frac{e_{th}^{\overline{AI}}}{1+e_{th}^{\overline{AI}}}$:

$$\Delta per_{th} = -l_{tc}^{AI}(1 - p_{t1h})\frac{e_{th}^{AI}}{1 + e_{th}^{AI}} + (1 - l_{tc}^{AI})p_{t1h}\frac{e_{th}^{\overline{AI}}}{1 + e_{th}^{\overline{AI}}} \quad (24)$$

Thus, the performance effect of withholding AI advice on a specific task is positive, if the likelihood of correct advice is below the ratio of harm of incorrect advice and the sum of the harm of incorrect and the benefit of correct advice.

$$l_{tc}^{AI} < \frac{p_{t1h}\frac{e_{th}^{\overline{AI}}}{1 + e_{th}^{\overline{AI}}}}{p_{t1h}\frac{e_{th}^{\overline{AI}}}{1 + e_{th}^{\overline{AI}}} + (1 - p_{t1h})\frac{e_{th}^{AI}}{1 + e_{th}^{AI}}} = r_h \quad (25)$$

We denote r_h as the *critical ratio* of an individual human.

Effect on Human Performance

Let $\mathcal{T}(l_{tc}^{AI} < r_h)$ be the set of tasks, where the AI certainty $l_{tc}^{AI} < r_h$. We denote $\Delta UHK_h = \frac{1}{T}\sum_{t\in\mathcal{T}^{\overline{AI}}(l_{tc}^{AI}<r_h)} p_{t1h}$ as the "conserved" unique human knowledge due to not receiving AI advice, and $\Delta UAK_h = \frac{1}{T}\sum_{t\in\mathcal{T}^{AI}(l_{tc}^{AI}<r_h)}(1 - p_{t1h})$ as the potentially "lost" unique AI knowledge due to withholding AI advice. The ratio $\frac{\Delta UHK_h}{\Delta UAK_h}$ is denoted as change in relative unique human knowledge. We can derive the total effect of personalizing AI advice for a human h over a set of tasks $t \in \mathcal{T}$ as follows:

$$\Delta per_h = \Delta UHK_h\frac{e_h^{\overline{AI}}}{1 + e_h^{\overline{AI}}} - \Delta UAK_h\frac{e_h^{AI}}{1 + e_h^{AI}} \quad (26)$$

The net effect of personalizing advice can be expressed as the difference between expected mitigation of reduction of unique human knowledge due to incorrect AI advice (with $l_{tc}^{AI} < r_h$) and the missed benefit of unique AI knowledge due to correct AI advice (with $l_{tc}^{AI} < r_h$).

$$\delta_{eh} < \frac{\Delta UHK_h}{\Delta UAK_h} \Rightarrow \Delta per_h > 0 \quad (27)$$

We can now derive our third theoretical proposition.

Proposition 3. The overall performance effect of personalizing AI advice is positive in the following case: The scaling factor of the effort of correct advice (δ_{eh}) is smaller than the change in relative unique human knowledge.

Interestingly, human ability to effectively differentiate between correct and incorrect advice δ_{eh} decreases the potential benefit of personalized AI advice, whereas the ability of an AI to differentiate between tasks with unique AI knowledge and unique human knowledge (i.e., change in relative unique human knowledge) increases the potential benefit of personalized AI advice.

We can now specify our next hypothesis, conservatively, regarding the effect of personalized advice on human performance:

Hypothesis 3a: Human accuracy does not decrease when AI advice is personalized.

Effect on Unique Human Knowledge

We denote the effect of providing personalized AI advice on unique human knowledge as ΔUHK_h^{per}:

$$\Delta UHK_h^{per} = \frac{e_h^{\overline{AI}}}{1 + e_h^{\overline{AI}}}\Delta UHK_h \quad (28)$$

Thus, as long as the AI withholds any incorrect advice, and humans put any weight on incorrect advice, providing AI advice individually has a positive effect on unique human knowledge. Consequently, we can specify our final hypothesis regarding the impact of personalized advice on unique human knowledge:

Hypothesis 3b: Unique human knowledge increases when AI advice is personalized.

In summary, our theoretical framework predicts that AI advice increases human performance, but decreases unique

human knowledge. Our two interventions, at least, retain the benefits of AI advice on individual performance (i.e., individual performance does not reduce in comparison to a case in which there is no intervention) while mitigating the loss in unique human knowledge.

Theoretical Framework: Wisdom of Crowds

We discuss wisdom of crowds setting to illustrate the potential effect of AI advice on environments where decisions are based on multiple human opinions. In a wisdom of crowds setting, a group of n humans solve tasks t, that is, to select the correct choice out of C possible choices. We denote the number of humans selecting choice c as n_c. The modal choice, that is, $\max(n_c, c \in \mathcal{C} = \{1 \dots C\})$, is defined to be the group choice.

Illustrative Example. To motivate the effect of AI advice on wisdom of crowds settings, we start with a simple illustrative example of two choices and three individuals (i.e., $C = 2$, $n = 3$), and a set of homogenous tasks $t \in \mathcal{T}$ with $p_{t1} = 0.7$ for all t and $\frac{1}{T}\sum_{t\in\mathcal{T}} a_{t1} = 0.75$, that is, an average human accuracy of 0.7 and an average AI accuracy of 0.75. For this example, let us assume the extreme case: that humans always follow AI advice and, therefore, improve individually by five percentage points to an accuracy of 0.75. However, if all humans follow the AI signal, they are all correct or all incorrect for the same task, making the AI choice the modal choice for each task. Consequently, the group performance equals 0.75 for all possible group sizes. Now consider a case in which there is no AI advice. In this case, humans decide independently of one another and the group results follow a binomial distribution with a probability distribution of $P_t(n_{t1}, n) = \frac{n!}{n_{t1}!(n-n_{t1})!} p_{t1}^{n_{t1}}(1-p_{t1})^{n-n_{t1}}$; then, the resulting group accuracy is $\sum_{n_{t1}=\left\lfloor\frac{n-1}{2}\right\rfloor+1}^{n} P(n_{t1}, n)$. For our example, the group decides correctly if two or three of the group members select the correct choice, that is,

$$P_t(2,3) + P_t(3,3) = \frac{3!}{2!\,1!}0.7^2 0.3^1 + \frac{3!}{3!\,0!}0.7^3 0.3^0 = 0.784 > 0.75.$$

While this example is illustrative, we can show that for any case where the correct choice is more likely to be selected than the incorrect choice ($p_{t1} > 0.5$), the group accuracy will exceed any potential AI performance (< 1). The probability of an incorrect group decision equals the cumulative binomial distribution function $\sum_{n_{t1}=0}^{0.5\cdot n} P_t(n_{t1}, n)$. As no closed-form solution exists, we make use of Hoeffding's inequality (Equation (29), Hoeffding 1963), that provides an upper bound for the cumulative distribution function for $k < np$ (k being number of successes). Thus, for $p > 0.5$ and $k = 0.5 \cdot n$ we can derive the following:

$$F(k, n, p) \leq \exp\left(-2n\left(p - \frac{k}{n}\right)^2\right) \tag{29}$$

$$\sum_{n_{t1}=0}^{0.5\cdot n} P_t(n_{t1}, n) \leq \exp(-2n(p_{t1} - 0.5)^2) \tag{30}$$

As $(p_{t1} - 0.5)^2 > 0$, the right-hand side of (30) is decreasing in n and converging toward zero. We may derive the fourth proposition:

Proposition 4. For settings with two choices and a higher probability to select the correct choice, a group size exists in which the group accuracy exceeds any threshold accuracy value below one.

Consequently, the probability that the alternative choice is the group choice decreases with increasing group size and converges toward zero. Note that our model is much more general and subsumes this example as one of the possible environments. In the following, we define the condition for a much more general setting where individual accuracies are heterogeneous and humans may or may not follow the AI advice even if it has a higher probability of being correct and AI advice simply changes the probability of selecting a choice with effect strength e_t as discussed earlier in our model development.

Group Accuracy with and Without AI Advice

We can model the probability of a given outcome with a group of individuals using a multinomial distribution as follows:

$$P_t(n_{t1}, n_{t2}, \dots, n_{tC}) = \frac{n!}{\prod_{c=1}^{C} n_{tc}!} \prod_{c=1}^{C} {p_{tc}}^{n_{tc}} \tag{31}$$

The probability that a group of size n selects choice c for task t is denoted as P_{tc}^n

$$P_{tc}^n = P_t(n_{t1}, n_{t2}, \dots, n_{tC} | n_{tc} \geq n_{ti}, i \in \{1..C\}\backslash c) \tag{32}$$

Our tie-breaking rule assumes that ties are broken randomly. Accordingly, we compute P_{t1}^n as follows:

$$P_{t1}^n = \sum_{n_{t1}=\left\lceil\frac{n-1}{C}\right\rceil}^{n} \sum_{i=2}^{C} \sum_{n_{ti}=LB_{ti}}^{UB_{ti}} \frac{1}{\left|\{n_{tj}, n_{tj} = n_{t1}\}\right|} P(n_{t1}, n_{t2}, \dots, n_{tC}) \tag{33}$$

$$LB_{ti} = \max\left(0, n - \sum_{j=1}^{i-1} n_{tj}\right) - (C-i)(n_{t1})$$

$$UB_{ti} = \min\left(n_{t1}, n - \sum_{j=1}^{i-1} n_{tj}\right)$$

The first sum considers the number of humans selecting the correct choice, where the lower bound ensures that outcomes exist where no other choice is selected more often. The second sum considers all other choices i, and the third sum considers the number of humans selecting choice i. The lower bound ensures that no choice needs to be selected more often than the correct choice: $n - \sum_{j=1}^{i-1} n_{tj}$ is the number of humans who haven't selected a choice yet and $(C-i)(n_{t1})$ is the maximum number of humans selecting the succeeding choices without exceeding the correct choice. The upper bound is the number of humans selecting the correct choice or the number of humans who haven't selected a choice yet. For example, having $n = 3$ humans and $C = 3$ choices would lead to $P_{t1}^3 = \frac{1}{3}P(1,1,1) + \frac{1}{1}P(2,0,1) + \frac{1}{1}P(2,1,0) + \frac{1}{1}P(3,0,0)$. Note that for $n \leq 3$, $P_{t1}^n = p_{t1}$.

Effect of AI Advice on Wisdom of Crowds' Performance

As discussed earlier, AI advice changes the probability of selecting choice c in task t with effect strength e_t. We now adapt the probability of a given outcome of Equation (31) for correct AI advice (34) and incorrect AI advice (35). Note that for simplicity, we define the incorrect AI choice to be $c \in \{2\}$, that is, $a_{t2} = 1$.

$$P_t^{AI}(n_{t1}, n_{t2}, \ldots, n_{tC}) = \frac{n!}{\prod_{c=1}^{C} n_{tc}!} \frac{p_{t1} + e_t}{1 + e_t}^{n_{t1}} \prod_{c=2}^{C} \frac{p_{tc}}{1 + e_t}^{n_{tc}} \quad (34)$$

$$P_t^{\overline{AI}}(n_{t1}, n_{t2}, \ldots, n_{tC}) = \frac{n!}{\prod_{c=1}^{C} n_{tc}!} \frac{p_{t1}}{1 + e_t}^{n_{t1}} \frac{p_{t2} + e_t}{1 + e_t}^{n_{t2}} \prod_{c=3}^{C} \frac{p_{tc}}{1 + e_t}^{n_{tc}} \quad (35)$$

Given the generalized nature of this formulation, it is not possible to obtain a closed-form analytical result regarding the size of a crowd without AI assistance that outperforms a crowd with AI assistance. However, the equations (34) and (35) easily lend themselves to exploration of the structure of results for any practical situation where the required probabilities can be calculated or empirically obtained.

For example, we can compute the crowd performances P_{t1}^n for our experimental environment with 10 choices (i.e., $C = 10$), assuming an effect size of $e = 0.5$ and equal probabilities for all non-correct choices with $p_j = \frac{1-p_1}{9}$. Let us compare three different tasks with different human accuracy: a "difficult" task with $p_1 = 0.2$, a "medium" task with $p_1 = 0.5$, and an "easy" task with $p_1 = 0.8$. We vary the group size between 1 and 25, and illustrate the results in Figure 1.

Note that with increasing group size, the probability that the crowd selects the choice with the highest probability increases. Thus, correct advice helps crowds to converge faster. However, after a certain group size, the effect is negligible. Incorrect advice slows down convergence (if the correct task still has the highest probability, as may be the case for easy tasks), or even turns convergence toward a performance of zero (if the correct task no longer has the highest probability, as may be the case for difficult tasks).

For our examples, incorrect advice has a strong detrimental effect on group performance for all task types, while correct advice only has a strong beneficial effect for the difficult task.

The performance of groups with AI advice depends on the ratio of correct and incorrect advice, that is, AI accuracy. With increasing group size, the relative benefit from correct advice deteriorates for cases where the initial probability for the correct choice is the highest. In Figure 2, we illustrate the results for human accuracies between 0.1 and 0.9 and AI accuracies between 0.1 and 0.9. The values in the matrix represent the group size at which performance of groups without AI advice begins to outperform groups with AI advice (light grey area reflects that a single human outperforms a human with AI advice). Human accuracy of 0.1 illustrates a special case, where all choices have the same probability to be selected. If AI accuracy equals 0.1 as well, groups with and without AI advice perform equivalently for all group sizes (white area), while groups with AI advice outperform those without for all group sizes (dark grey area) if AI accuracy is greater than 0.1 and human accuracy is 0.1. In general, as group size grows, humans without AI assistance will at some point outperform a same-sized group with AI assistance. For example, if human accuracy is 0.5, and AI accuracy is 0.9, a group of 11 or more humans will perform better without AI advice as compared to the same number of humans with AI advice.

Thus, for all situations where the correct choice has the highest probability of being chosen, there is a group size where groups without AI advice outperform those with AI advice, even for higher AI accuracies.

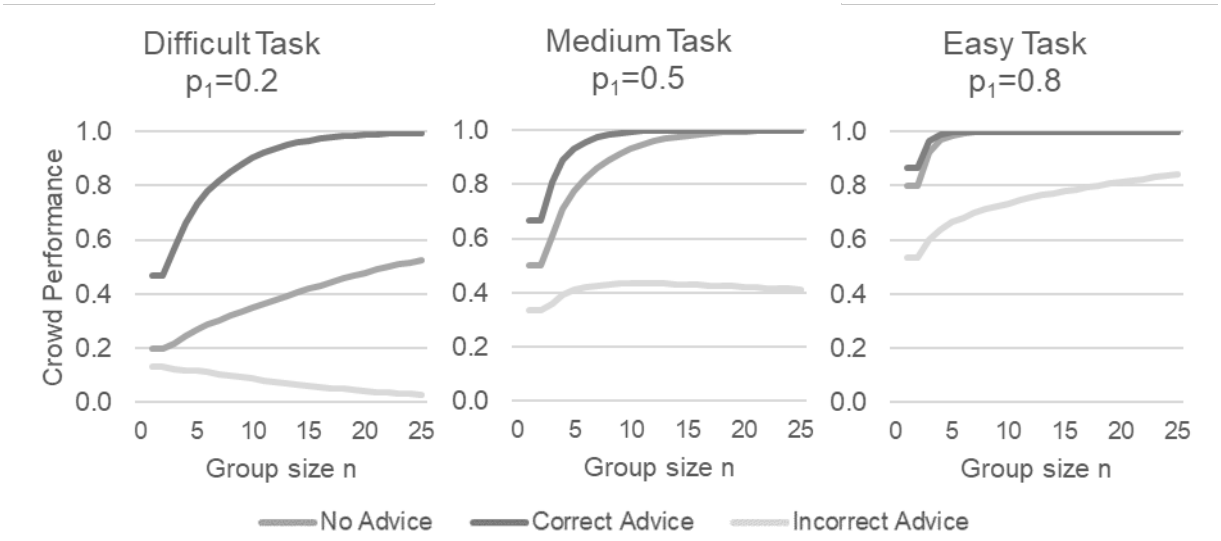

Figure 1. Group Performance Based on Human Accuracy, Advice, and Group Size

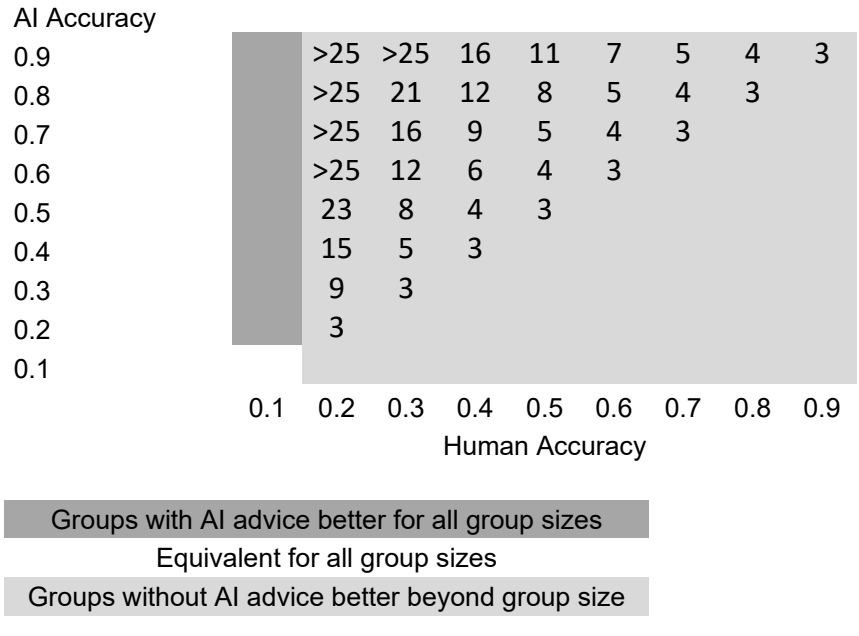

Figure 2. Comparison of Groups with and Without AI Advice Based on Human and AI Accuracy

The diversity prediction theorem (Page 2007) states that collective performance depends on individual performance and prediction diversity. This statement alone makes it difficult to predict how AI suggestions will impact crowd performance because while the AI advice increases individual performance, it decreases unique human knowledge and, thus prediction diversity. In line with our model results, Hong et al. (2016) argue that in general, the importance of diversity increases with group size. Considering those theoretical considerations in line with Proposition 4 and the numerical results, we derive our next hypothesis.

Hypothesis 4a: Benefit from wisdom of crowds decreases with the group size when receiving AI advice.

The final hypotheses directly follow Page's diversity prediction theorem and our hypotheses on individual decision making. Compared to providing regular AI advice, both our interventions should at least maintain individual performance while increasing unique human knowledge as measure for prediction diversity.

Hypothesis 4b: Benefits from wisdom of crowds increase with the group size when receiving the AI's certainty.

Hypothesis 4c: Benefits from wisdom of crowds increase with the group size when AI advice is personalized.

Next, we present our experiments to study individual decision making and to test the hypotheses derived from our theoretical model, before performing computational experiments from the data collected in the individual decision-making environments to simulate wisdom of crowds settings to validate the predictions of our model.

Experimental Studies: Individual Decision Making

To empirically test our hypotheses, we conducted a set of experimental studies with human subjects. We address the questions whether humans can exploit complementarities with an AI, and how the level of unique human knowledge changes due to different types of advice. We provide an overview on our experiments and the high-level purpose of each experiment in Table 1. In Experiment 1, we initially test the consequences of AI advice and test the effect of presenting AI certainty along with AI suggestions. In Experiment 2, we explore the heterogeneity in human reactions to advice. Finally, Experiment 3 focuses on the effect of personalized AI advice.

We chose the context of image recognition for three main reasons: First, image recognition is a very generic task that all human subjects should be able to perform without any specific skills or training. In behavioral research, a common goal is to create a setting where insights also apply in other settings. It is assumed that observations in generic tasks can carry over to more specialized tasks, while contexts that do require specific training make results less generalizable. Second, image classification is a task that modern AI systems do well (Szegedy et al. 2015) and at least on par with human performance (Russakovsky et al. 2015). Third, image classification is a task where an AI based on a deep convolutional neural network will use different solution methods to classify an image compared to any individual human. Thus, the AI should perform better than a given human on some images, but worse on other images, leading to the existence of unique AI knowledge and unique human knowledge.

General Experimental Design

In the experiment, the subjects had to assign a focal image (e.g., an image of a small black dog) to one of ten possible image classes. For each of the 10 classes, we showed the class name (for example the text "Swiss mountain dog" or "Boxer") and 13 images that belong to that class, similar to Russakovsky et al. (2015). We sampled 100 images and the corresponding correct class labels from the ImageNet database (www.image-net.org). All subjects classified the same 100 focal images. Images contain different levels of subjective difficulty; for example, humans may find classifying a firetruck more straightforward than identifying a specific breed of dog. After each classification, subjects reported how certain they were about their choice on a four-point scale ("Uncertain 1/4," "Rather uncertain 2/4," "Rather certain 3/4," "Certain 4/4").

All studies include two main conditions: The subjects in the "No AI" conditions worked alone. They made the decisions by themselves and did not receive any help. In the "AI suggestion" conditions, subjects received advice from one of the best performing AIs for image classification—GoogLeNet Inception v3 (Szegedy et al. 2016). For each image, Inception assigns a certainty score to 1,000 possible classes. This score represents the likelihood that the selected class is the true class for a given focal image. The AI, in our experiment, recommends the class with the highest certainty score to the subjects. In our set of 100 images, the AI had an accuracy of 0.77, that is, it classified 77 of the 100 images correctly. Figures 3 and 4 show screenshots of those treatments. They illustrate that the only difference between them is the AI's suggestions.

Table 1. Overview Experiments

Experiment	Purpose
Experiment 1: AI advice and the effect of providing AI certainty	• Effect of AI advice on human accuracy and unique human knowledge • Effect of providing the AI's certainty on human accuracy and unique human knowledge
Experiment 2: Heterogeneity in human reactions to AI advice	• Effect of correct and incorrect AI advice on individual human accuracy (benefit and harm)
Experiment 3: Personalizing AI advice	• Effect of personalizing AI advice on human accuracy and unique human knowledge

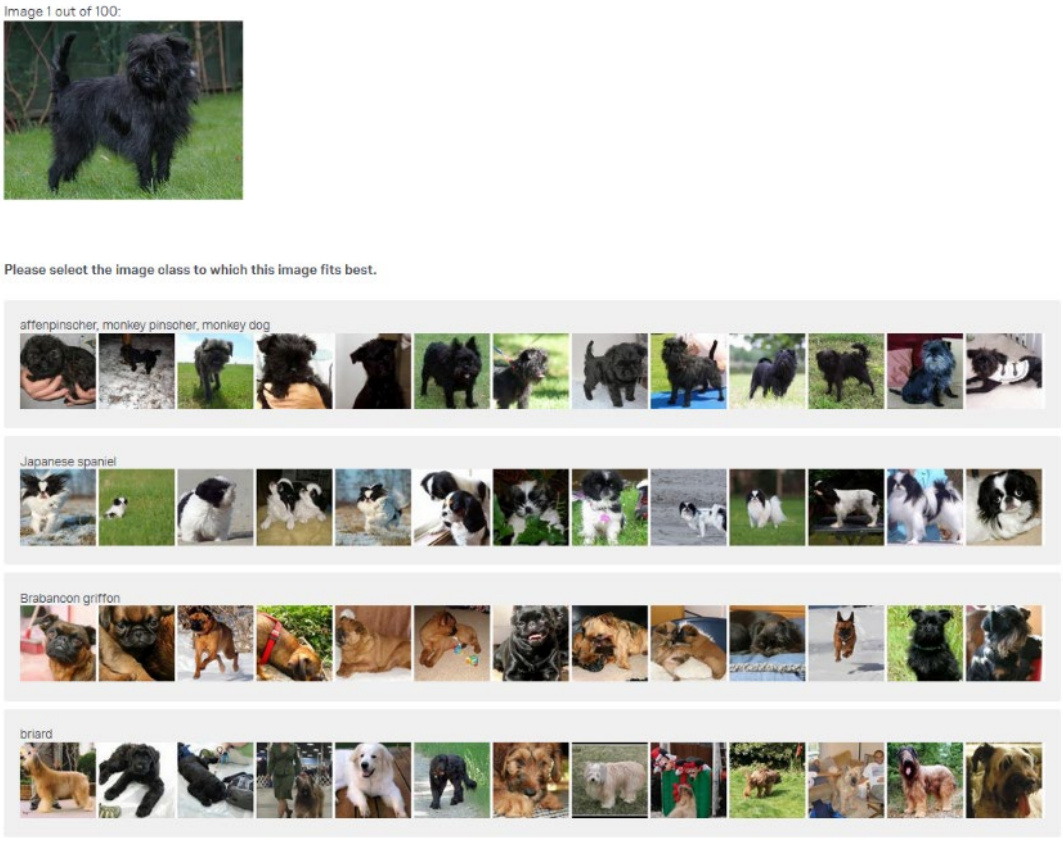

Figure 3. Humans Without AI Advice (only four out of ten choices are shown)

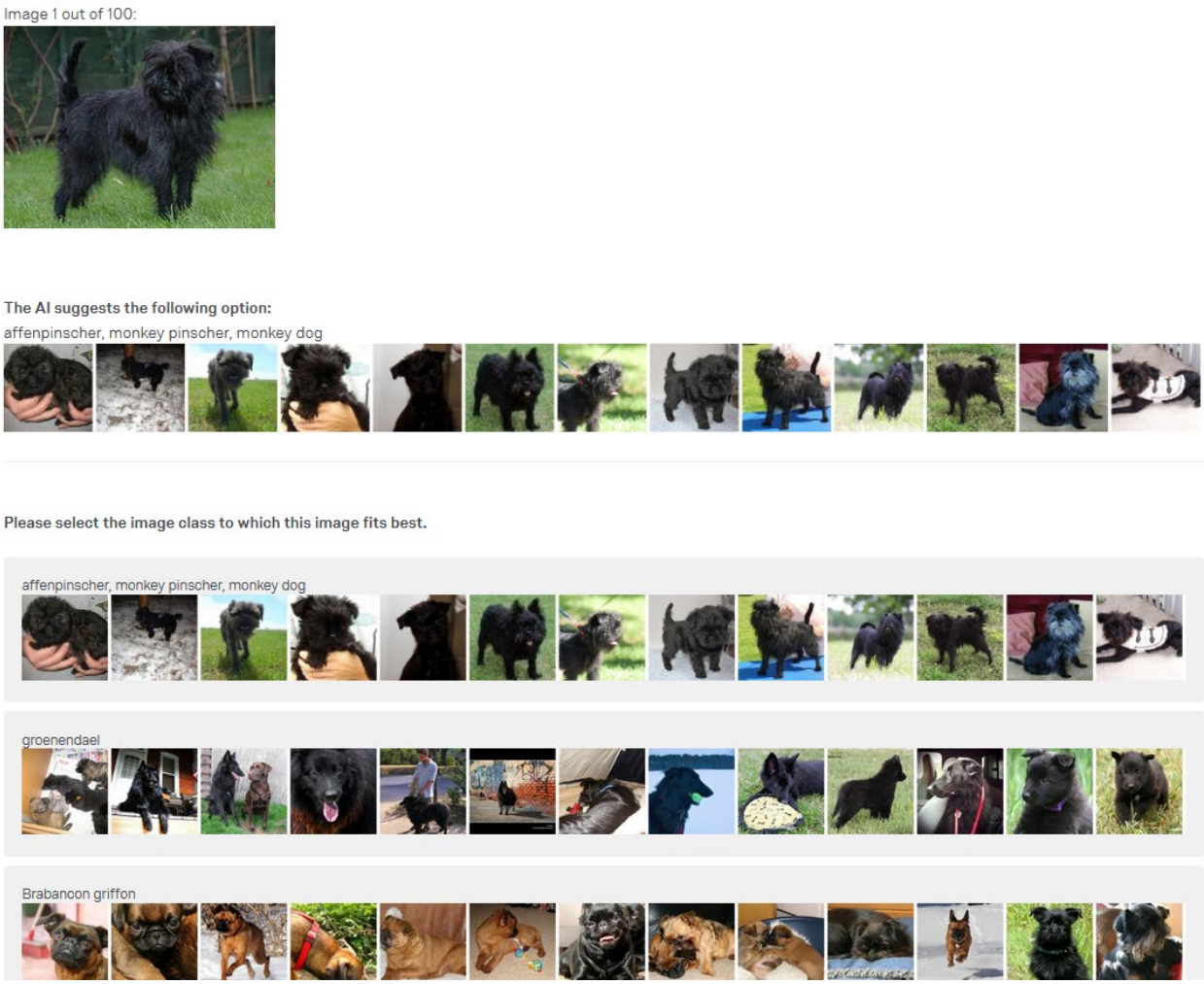

Figure 4. With AI Suggestions (only four out of ten choices are shown)

We followed Nosek et al. (2018) and pre-registered the two confirmatory Experiments 1 and 3 (https://osf.io/b6se4/) at the Open Science Foundation (Foster and Deardorff 2017). This included our hypotheses, a power analysis to determine the sample size, the recruitment and data collection process, and the statistical analysis. Doing experimental work this way ensures that research is driven by theory and the reported results are not an outcome of *ex post* analysis.

Measures

As in our theoretical model, we are interested in two measures: human accuracy and unique human knowledge. Human accuracy is measured as the number of correctly classified images divided by the total number of images. Unique human knowledge is measured as the number of correctly classified images that the AI classified incorrectly,

divided by the number of images. We hypothesize, test and report based on both measures in the experiments.

In the following, we describe the three experiments.

Experiment 1: AI Advice and the Effect of Providing AI Certainty

Hypotheses and Study Design

We set up Experiment 1 to test the initial hypotheses on the consequences of AI advice on human accuracy and unique human knowledge (Hypotheses 1a and 1b), and the effect of presenting AI certainty (Hypotheses 2a and 2b). We compare three experimental treatments in a fully randomized between-subjects design: Treatment 1 "No AI" and Treatment 2 "AI suggestion" as described above, and an additional Treatment 3 "AI certainty," where advice additionally includes information about the AI's certainty. To ensure human subjects can easily understand the AI certainty and can relate it to their own level of certainty, we presented the AI certainty on the same four-point scale that the human used to report their own certainty (subjects were informed about the likelihoods of the categorical certainty scores and how they relate to average human performance).

We summarize the four hypotheses we test in Experiment 1 and how they relate to our treatments:

- Hypothesis 1a: Human accuracy increases when receiving AI advice (accuracy in Treatment 1 is lower than accuracy in Treatment 2).
- Hypothesis 1b: Unique human knowledge decreases when receiving AI advice (unique human knowledge in Treatment 1 is greater than unique human knowledge in Treatment 2).
- Hypothesis 2a: Human accuracy increases when receiving the AI's certainty (accuracy in Treatment 2 is lower than accuracy in Treatment 3).
- Hypothesis 2b: Unique human knowledge increases when receiving the AI's certainty (unique human knowledge in Treatment 2 is lower than unique human knowledge in Treatment 3).

Study Protocol

We performed the experiment on August 8, 2019. We preregistered a power analysis to determine our sample size. We decide to test for a small to medium-sized effect ($d = 0.3$, $f = 0.15$) assuming an alpha value of 0.05 and a power of 0.8 (ANOVA with three groups). This results in a total sample size of 432. We targeted 150 per cell resulting in a total of 450 subjects. We recruited 458 subjects on Amazon Mechanical Turk (MTurk). We only included subjects from the United States who had a positive rating of at least 90%, had not participated in related studies before, correctly answered an attention check, and met technical requirements regarding screen resolution.

Subjects received $1 for participation, $1 for correctly estimating the number of classified images (+/– five images) after the 100 classifications, and $0.05 for each correctly classified image. Subjects in Treatments 2 and 3 received an additional $0.50 for filling out a survey about their trust in the AI after the classification. Total payment could thus vary between $1 and $7 for Treatment 1, and between $1.5 and $7.5 for Treatments 2 and 3.

Random assignment put 146 subjects into Treatment 1, 160 subjects into Treatment 2, and 152 subjects into Treatment 3. We collected information on the subjects' gender, age, level of education, and income class. We control for these factors in robustness checks to make sure that the randomization did not systematically influence treatment effects.

At the beginning of the experiment, subjects received basic information on the task and had to perform an attendance check. We informed them that they could only continue if all answers were correct. Afterward, they were randomly assigned to an experimental treatment and received instructions that only differed with respect to the advice provided. In the main task, subjects classified 100 images in random order. The possible classes for each image were presented in random order. After the classifications, subjects had to estimate the number of correctly classified images. They reported how they made the decisions, with the subjects of treatments with AI advice (Treatments 2 and 3) also answering a questionnaire on human–computer trust covering the dimensions perceived reliability, perceived technical competence, understandability, faith, and personal attachment (10 items from Madsen and Gregor 2000). The experiment ended with a short demographic questionnaire. Subjects were then told the results and their payment. Average duration was 57.4 minutes, average pay excluding Amazon MTurk fees was $5.77.

Results

We first test our pre-registered hypotheses[2] on the effect of AI advice and providing AI certainty on human accuracy and unique human knowledge. Summary statistics for accuracy and unique human knowledge can be found in Table 2, the mean outcomes are illustrated in Figure 5.

We present results on accuracy (Hypotheses 1a and 2a) and the interpretation of effect sizes (small, medium, large) according to Cohen (2013). The variance of accuracy is significantly different across experimental conditions (Levene test, $F(2, 455) = 17.477$, $p < .001$) and the means are significantly different as well (ANOVA with heterogeneous variances, $F(2, 285.29) = 29.594$, $p < .001$, $\eta^2 = .155$, which represents a large effect). *Post hoc* comparisons with Tanhames T2 statistic for multiple comparisons suggest that accuracy with AI suggestion (Treatment 2, 0.799) is indeed significantly larger than without (Treatment 1, 0.681). This difference (11.8 percentage points) is significant ($p < .001$) and represents a large effect ($d = .85$). Additionally, showing the AI's certainty in Treatment 3 created an average accuracy of 0.801. The difference between Treatments 2 and 3 is not significant ($p > .99$) and would represent not even a small effect ($d = .02$).

Thus, we find support for Hypothesis 1a: accuracy improves when humans are provided AI suggestions. In addition to our pre-registered hypothesis we also conclude that human accuracy (0.799) exceeds the AI accuracy level of 0.770 when AI suggestion is provided ($p < .001$). We do not find sufficient support for Hypothesis 2a, and cannot conclude that accuracy improves further when the AI provides information about its certainty.

The variances of unique human knowledge across treatments does not vary significantly according to a Levene test ($F(2,455) = 1.807$, $p = .165$). Analysis of variance indicates that the mean unique human knowledge differs across treatments ($F(2,455) = 47.336$, $p < .001$, $\eta^2 = .172$, which represents a large effect). Average unique human knowledge is .123 when humans are working alone, and AI suggestion reduces the unique human knowledge down to .073. Tukey's honestly significant differences (HSD) indicates that this difference of five percentage points is significant ($p < .001$), and represents a large effect ($d = 1.125$). When we additionally present the AI's certainty (Treatment 3), the unique human knowledge is .087. This is significantly different from both other treatments ($p < .001$ in both cases). The difference between Treatments 2 and 3 is 1.4 percentage points, which represents a small effect ($d = .285$). Therefore, we find support for Hypotheses 1b and 2b, that is, unique human knowledge is reduced with AI suggestion, and providing information about the AI's certainty partially mitigates this effect.

As discussed earlier, Hypothesis 2a that posited a further increase in human accuracy when AI certainty is provided, was not supported. Our prediction based on our model was that humans would follow advice less often for images with low AI certainty and follow the AI's advice more often when the AI is certain. Since the AI advice is more likely to be correct for high certainty values, hence, we expected an increase in accuracy. However, that did not turn out to be the case.

To explore this negative result further, we measure the effect of providing AI certainty on human accuracy on image level and differentiate between images where the AI advice was correct and where it was wrong. We further consider image difficulty as control variable, and differentiate between image difficulty from a human point of view (1 minus human accuracy in T1) and from an AI point of view (1 minus AI certainty score). We run three regression models (Table 3) with human accuracy as the dependent variable. Model (1) was estimated without controls for image difficulty. Models (2) and (3) include controls for image difficulty from a human (model 2) and the AI (model 3) perspective. We use random effects models to capture image and subject heterogeneity. The results suggest that providing information about the AI's certainty increases human accuracy by 0.059 for images where AI was wrong. However, we also see a negative interaction effect of providing AI certainty for images where the AI is correct: average accuracy actually declines for images where the AI's suggestion was correct. Thus, providing AI certainty led to a consistent decrease of AI adherence. Note that, according to our theoretical model, this is potentially harmful if, as is the case in our experiments, AI accuracy exceeds human accuracy (and unique AI knowledge exceeds unique human knowledge).

In our survey questions on human–computer trust (Madsen and Gregor 2000), we observed that providing AI certainty led to a decrease of human trust in AI's capabilities. As trust in AI is typically correlated with adherence to advice (see Glikson and Woolley 2020 for a review on trust in AI), the

[2] Note that we preregistered hypotheses based on human–human complementarities as well. As we focus on unique human knowledge (pre-registered as human–AI–complementarities), we skip those hypotheses.

Table 2. Descriptive Statistics for Accuracy and Human–AI–Complementarity

	Treatment	n	Minimum	Average	Median	Maximum	Std. Dev.
Accuracy							
	1	146	.08	.681	.735	.89	.172
	2	160	.06	.799	.82	.92	.098
	3	152	.4	.801	.84	.94	.109
Unique human knowledge							
	1	146	.01	.123	.13	.19	.043
	2	160	0	.073	.08	.17	.046
	3	152	0	.087	.10	.19	.049

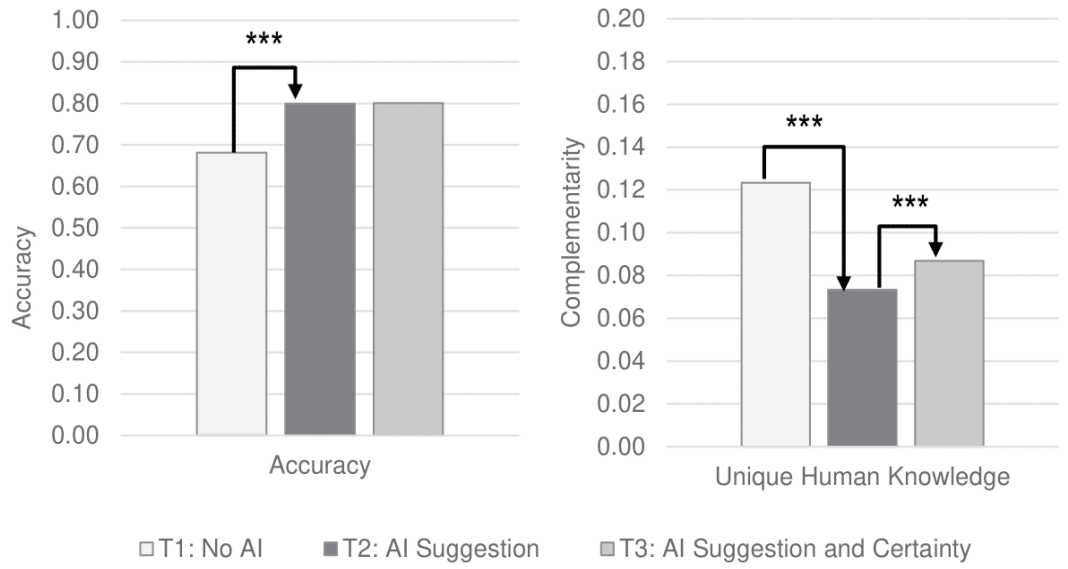

Figure 5. Average Values for Accuracy and Unique Human Knowledge across Treatments (*$p < 0.001$)**

Table 3. The Effect of Providing AI Certainty (T3) on Human Accuracy (1), Considering Image Difficulty from a Human Point of View (2), and from an AI Point of View (3)

	Accuracy (1)	Accuracy (2)	Accuracy (3)
Constant (T2, AI wrong)	.318***	.505***	.328***
AI certainty (T3)	.059***	.059***	.059***
Human difficulty		-.402***	
AI difficulty			-.018(*)
AI correct	.624***	.548***	.617***
AI certainty (T3) x AI correct	-.074***	-.074***	-.074***
Adjusted R^2	.408	.453	.408

Significance values: ***p < 0.001, **p < 0.01, *p < 0.05 (*)p < 0.1

decreased adherence to AI advice when AI certainty was provided might be explained by the observed decrease in trust.

When we compare mean accuracies from Treatment 1 and Treatment 2 in our first experimental study, we observe a gain of about 12 percentage points on average. While the AI advice boosts accuracy from 0.725 to 0.942 when it is correct, it deteriorates performance from 0.536 to 0.318 if it is incorrect. The former effect measures the "benefit" of correct advice, and the latter the "harm" of incorrect advice. However, individually, humans may react differently to advice: decisions of some humans might be robust against incorrect advice if they have better judgement for a given image, while other individuals might adopt the AI suggestion even if they themselves would have chosen the correct answer but were not confident about their own choice. Similarly, some subjects might be willing to accept correct advice if it is in conflict with their own judgement, while others will stick to their incorrect judgement. Therefore, we designed our next experiment to explore this phenomenon.

Experiment 2: Heterogeneity in Human Reactions to AI Advice

Study Design

The goal of this experiment is to explore heterogeneity in human reactions to AI advice, specifically expected benefit and harm. To do this, we test "No AI" and "AI Advice" in a within-subjects design, where each subject classifies each image twice: after a first classification without support, we show subjects the AI suggestion and they may reconsider their choice and switch to the AI suggestion. This experiment is exploratory in nature and we did not preregister hypotheses.

For each subject, we measure the accuracy before and after receiving AI advice, the difference between these accuracies, the average benefit of correct advice (that is, the number of images where the subject was wrong before and correct after receiving the advice divided by the number of images with correct AI advice), and the average harm of incorrect advice (that is, the number of images where the subject was correct before and wrong after receiving the advice divided by the number of images with incorrect AI advice).

Study Protocol

We performed the experiment on March 30, 2020. As we do not test any hypotheses, we did not preregister a power analysis to determine the sample size, and targeted 100 subjects. We recruited 99 subjects on Amazon MTurk. The selection criteria were same as those in Experiment 1. Subjects received a payment of $1.50 for participation, up to $1 for correctly estimating the number of classified images (+/- five images), $0.50 for filling out the AI trust survey, and $0.04 for each correct classification, both before and after receiving AI advice. Total payment could vary between $2 and $11. Average pay was $7.73 with an average duration of 79.7 minutes.

Results

Table 4 shows summary statistics. In general, we could replicate the main results from Experiment 1, the effect of receiving advice being somewhat higher. All subjects improved due to AI advice. There was considerable variation across individuals for the benefit of correct advice and the harm of incorrect advice.

To illustrate that this variation cannot be explained by different levels of AI adherence, we plot benefit and harm for all subjects in Figure 6. The raw data shows a lot of variety and no clear pattern. Thus, reactions to advice seem to differ across individuals.

We can draw two conclusions from this exploration:

(1) Overall, AI advice seems to be beneficial for humans. We did not observe even a single subject that lost accuracy due to the AI advice.

(2) There is substantial heterogeneity in how incorrect advice harms human decision making, and how correct advice is beneficial. Therefore, a promising approach may be to design a personalized AI advice system that adapts to an individual by considering their individual expected harm or benefit based on the AI's observation while working with that individual.

Table 4. Descriptive Statistics for Accuracy Without/With AI Advice, Benefit and Harm of AI Advice

	N	Minimum	Average	Median	Maximum	Std. Dev.
Accuracy without AI	99	.090	.607	.680	.850	.196
Accuracy with AI	99	.290	.765	.810	.920	.120
Δ accuracy with/without AI	99	.020	.158	.130	.540	.104
Benefit	99	.039	.243	.208	.740	.135
Harm	99	.000	.128	.130	.391	.081

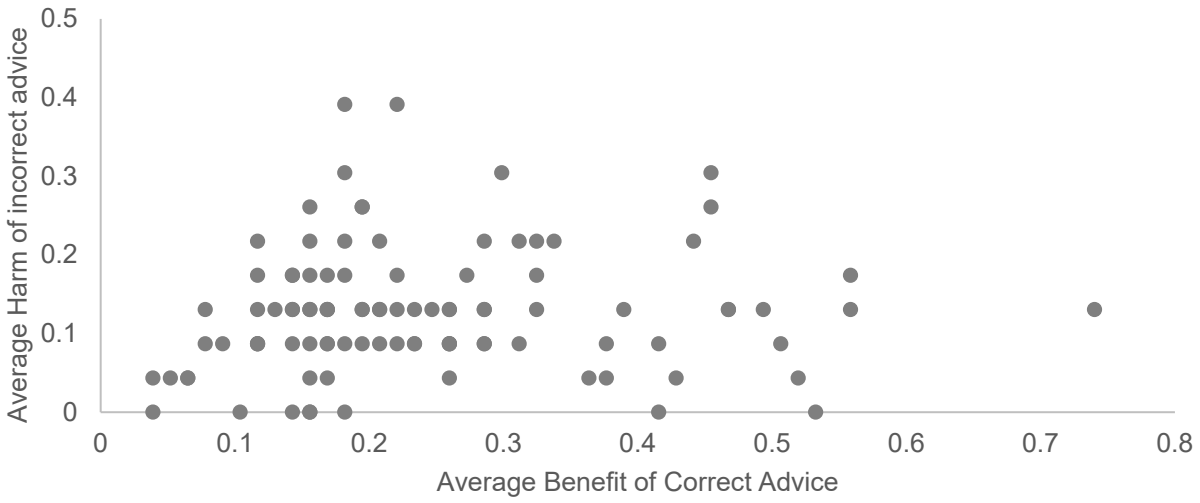

Figure 6. Average Benefit of Correct Advice and Harm of Incorrect Advice Per Subject

Experiment 3: Personalizing AI Advice

Hypotheses and Study Design

In Experiment 3, we test the remaining hypotheses on the effect of personalized AI advice on human accuracy and unique human knowledge. As derived in the theoretical model, in this case, the AI provides advice if the probability of correct advice at task t exceeds a "critical ratio" based on expected harm of incorrect advice and expected benefit of correct advice of human h:

$$l_t^{AI} < \frac{harm_h}{harm_h + benefit_h}$$

The experiment compares the personalized AI advice based on the critical ratio with two benchmark conditions "No AI" and "AI Suggestion" as in Experiment 1.

We split the 100 images in two sets of 50 images with similar attributes, such as average AI certainty, human performance, benefit of correct advice, and harm of incorrect advice. During the first set of 50 images, the AI "learns" about an individual, i.e., their expected benefit and harm when they receive AI advice. We test the three different experimental conditions during the second set of 50 images.

The experimental design of the first 50 images is equivalent to Experiment 2. We calculate the average benefit of correct advice, harm of incorrect advice, and the critical ratio for each subject. To exclude participants who randomly click, only subjects who classified more than 15 images correctly before receiving AI advice proceed to the second part. The

remaining subjects received their payment and left the study. Note that this process takes place before subjects are assigned to treatments to avoid any influence on the results. For the second 50 images, subjects were assigned to one of three treatments[3]: Treatment 1' ("No AI") and Treatment 2' ("AI suggestion") serve as benchmarks, and correspond to Treatments 1 and 2 in Experiment 1. Treatment 3' ("Personalized AI") is the new condition. Here we only provide a suggestion to a human if the AI certainty score exceeds the human's critical ratio. We pre-registered[4] the experiment at the Open Science Foundation following the similar protocol as Experiment 1.

Treatments 1' and 2' replicate our first experiment with a subset of images, so that we only hypothesize about effects between Treatments 2' and 3', that is, the effect of personalized advice. Based on our model, we restate our next set of hypotheses:

- Hypothesis 3a: Human accuracy does not decrease when AI advice is personalized (accuracy in Treatment 3' is not lower than in Treatment 2').

- Hypothesis 3b: Unique human knowledge increases when AI advice is personalized (unique human knowledge in Treatment 3' is greater than in Treatment 2').

Study Protocol

We performed the experiment on April 21, 2020. We preregistered a power analysis to determine our sample size: We decide to test for a small to medium-sized effect (d = 0.3, f = 0.15) assuming an alpha value of 0.05 and a power of 0.8 (ANOVA with three groups). This results in a total sample size of 432. We targeted 150 per cell resulting in a total of 450 subjects. After finishing the first set of 50 images, and after dismissing subjects with less than 16 images classified correctly, we randomly allocate subjects to treatments. As we expect that around 10% of subjects have to be dismissed after the first set of 50 images, we aimed for a total of 500 subjects. We recruited 492 subjects on Amazon MTurk. The selection criteria equaled those in Experiment 1. Subjects received a payment of $1.50 for participation, $1.50 for classifying for the main experiment, and $0.05 for each correct classification, both before and after receiving AI advice. Total payment could vary between $1.50 and $10.50. Average pay was $7.90 with an average duration of 68.6 minutes.

Results

We state summary statistics in Table 5 and illustrate the main outcomes in Figure 7. The variance of accuracy is significantly different across experimental conditions (Levene test, $F(2, 430)=7.7081$, $p < .001$) and the means are significantly different as well (ANOVA with heterogeneous variances, $F(2, 284.46) = 28.99$, $p < .001$, $\eta^2 = .127$, which represents a medium effect). *Post hoc* comparisons with Tanhames T2 statistic for multiple comparisons suggest that accuracy with AI suggestion (Treatment 2': .773) is indeed significantly larger than without (Treatment 1': .658). This difference (11.5 percentage points) is significant ($p < .001$) and represents a medium to large effect (d = .74). Showing personalized AI suggestions in Treatment 3' led to an average accuracy of .795. The difference between Treatments 3' and 2' is positive, yet not significant ($p > .1$); it would represent less than a small effect (d =.14). Thus, we find support for Hypothesis 3a: Accuracy does not decrease with personalized AI advice. We even find directional but nonsignificant support for better accuracy due to personalized advice.

The variance of unique human knowledge is not signifycantly different across experimental conditions (Levene test, $F(2, 430) = 0.117$, $p > .1$), but the means *are* significantly different (ANOVA, $F(2, 430) = 13.907$, $p < .001$, $\eta^2 = .06$, which represents a medium effect). *Post hoc* comparisons with Tukey's HSD suggest that unique human knowledge with permanent AI suggestions (Treatment 2', .074) is significantly smaller than without (Treatment 1': .108). This difference (3.4 percentage points) is significant ($p < .001$) and represents a medium effect (d = .62). The personalized suggestions in Treatment 3' created an average unique human knowledge of .097. Interestingly, the difference of 1.1 percentage points between unique human knowledge in Treatments 1' and 3' is not significant ($p > .1$). The difference between Treatments 2' and 3' (2.3 percentage points) is significant ($p < .01$) and represents a medium effect (d = .40). Thus, we also find support for Hypothesis 3b, as unique human knowledge increased significantly by

[3] In contrast to Experiment 1, we denote the treatments as 1' – 3'.

[4] Note that we preregistered hypotheses based on human–human complementarities as well. As we focus on unique human knowledge (pre-registered as "human–AI complementarities"), we skip those hypotheses.

Table 5. Descriptive Statistics for Accuracy and Human Knowledge

	Treatment	n	Minimum	Average	Median	Maximum	Std. Dev.
Accuracy							
	1’	139	.180	.658	.700	.960	.165
	2’	146	.060	.773	.800	.940	.144
	3’	148	.100	.795	.840	.960	.159
Unique human knowledge							
	1’	139	.000	.108	.120		.054
	2’	146	.000	.074	.080	.200	.056
	3’	148	.000	.097	.100	.220	.057

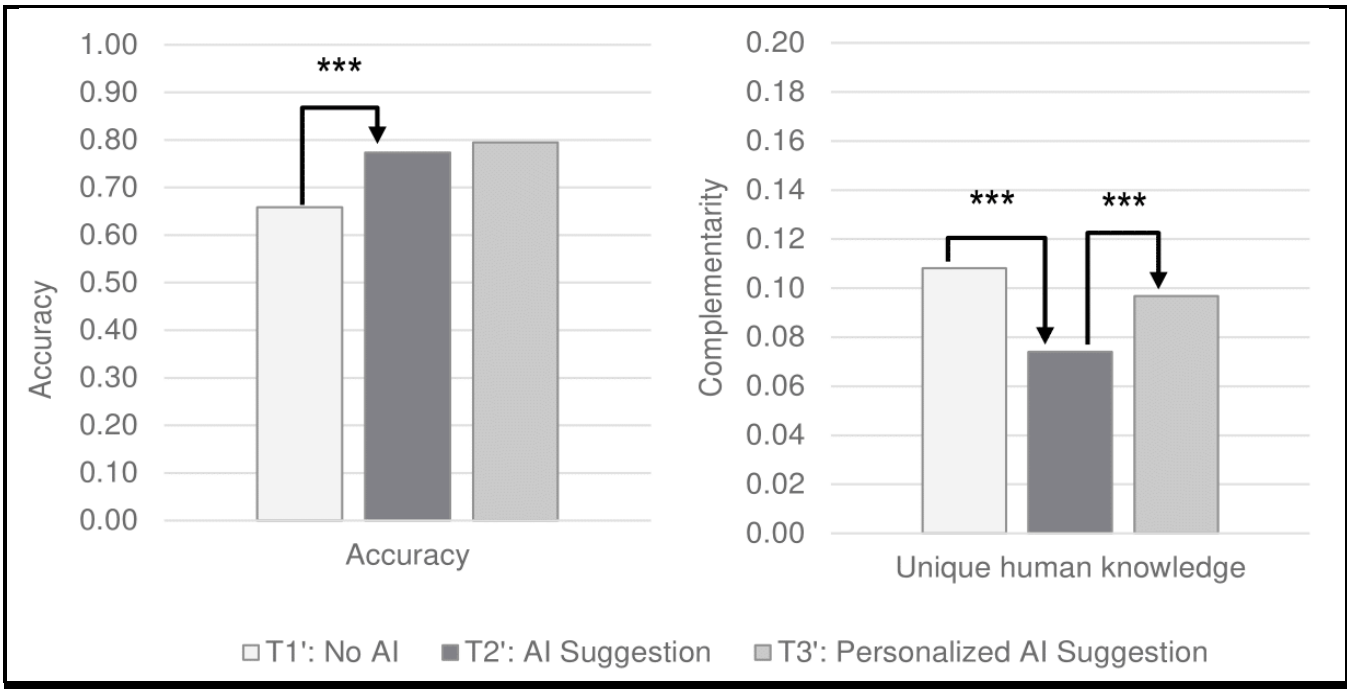

Figure 7. Average Values for Accuracy and Unique Human Knowledge Across Treatments (*$p < 0.001$)**

personalizing AI suggestions compared to always providing AI suggestions. In comparison to humans without AI advice, providing personalized suggestions significantly increased accuracy without significantly decreasing unique human knowledge.

The results for accuracy and unique human knowledge for the personalized AI resemble those of the third treatment from Experiment 1, where the AI certainty was provided. A difference between those approaches is that personalized AI prevents humans from converging toward false suggestions by not showing them.

We argue that the loss of diversity in human knowledge might have a negative impact in many collaborative decision environments. As a proof of concept, we next extend our theoretical model to incorporate wisdom of crowds settings, then we will use the data from our experimental studies to perform computational simulations to test the effects of AI advice and our two interventions in wisdom of crowd settings.

Experimental Studies: Wisdom of Crowds

In the earlier sections, our experimental results verified our model prediction that AI advice can increase human accuracy, but reduces unique human knowledge. We now study possible consequences of this negative effect and evaluate mitigation strategies based on wisdom of crowds. Using our theoretical model, we could demonstrate that following a common signal by AI advice leads to detrimental crowd performance. Even if AI accuracy was largely above average human accuracy, there is a group size where "pure" human groups outperform groups with AI advice—as long as the correct choice has a higher probability of being chosen over any other available choice. This finding is in line with the literature in group decision making and wisdom of crowds.

In the following, we make use of our experimental data to test our hypotheses.

Simulation Setup and Hypotheses

We simulated the wisdom of crowds, where the subjects of our different experimental treatments form "populations." For each crowd, we randomly select a desired number of group members. The collective choice is the modal choice among group members. In case of ties, we randomly select one of the tied choices. We vary the group size between one and 15. To explore convergence for large groups, we also include group sizes of 99 and 100. For each group size we perform a Monte Carlo simulation with 1,000 iterations. In each iteration we randomly sample the group from the population and compute its choices for all 100 images. We restate our set of hypotheses:

- Hypothesis 4a: Benefit from wisdom of crowds decreases with the group size when receiving AI advice.
- Hypothesis 4b: Benefit from wisdom of crowds increases with the group size when receiving the AI's certainty.
- Hypothesis 4c: Benefit from wisdom of crowds increases with the group size when AI advice is personalized.

In the following, we conduct this wisdom of crowds analysis for the experimental conditions of Experiment 1 and Experiment 3.

Wisdom of Crowds: Experiment 1

Figure 8 shows collective accuracies for populations without AI advice (Treatment 1), with AI suggestion (Treatment 2) and with AI suggestion and certainty (Treatment 3). The values for a group of one individual reflect the individual-level results from the previous section, where humans with AI suggestion (Treatments 2 and 3) outperform those without (Treatment 1). Due to the random tie-breaking rule, group sizes of one and two lead to the same performance.

The effect of AI advice on group performance changes quickly as the group becomes larger. The accuracy for humans with AI advice does not increase after reaching a certain group size: for AI advice without AI certainty, the accuracy reaches a level of .84 for group sizes above three, while it reaches a level of .86 for group sizes above four if AI certainty was added. Humans without AI advice keep improving over the entire spectrum. Starting with a group size of seven, human groups without AI advice outperform human groups with AI suggestions only. If AI suggestions are complemented with AI certainty, it takes groups of more than eleven humans without AI advice to outperform the groups with AI advice.

To test our Hypotheses 4a and 4b, we run a simple linear regression analysis with the group accuracy as the dependent variable. As independent variables we consider: group size (between 1 and 15), AI advice (1 for Treatments 2 and 3), the interaction of group size and AI advice, AI certainty (1 for Treatment 3), and the interaction of group size and AI certainty. For each treatment and group size, we consider the first 100 simulation instances. The regression results are summarized in the left column of Table 6. We obtain a significant negative interaction effect of AI advice and crowd size. Thus, humans with AI advice benefit less from crowd size compared to humans without AI advice, and we find support for Hypothesis 4a. Hypothesis 4b is also supported, as we find a significant positive interaction effect of AI certainty and crowd size. Thus, providing the AI's certainty mitigates the loss of crowd benefit of AI advice.

The results demonstrate that unique human knowledge is indeed essential. Although AI advice helps human decision makers to improve individually, the loss of unique human knowledge cancels out this improvement and ultimately leads to a deterioration of group performance. We identify two main reasons for this effect. First, the value of correct advice decreases by increasing group size, as groups are likely to select the correct decision without advice as well. Second, incorrect advice harms groups in two ways. Not only do humans who would have otherwise chosen the cor-

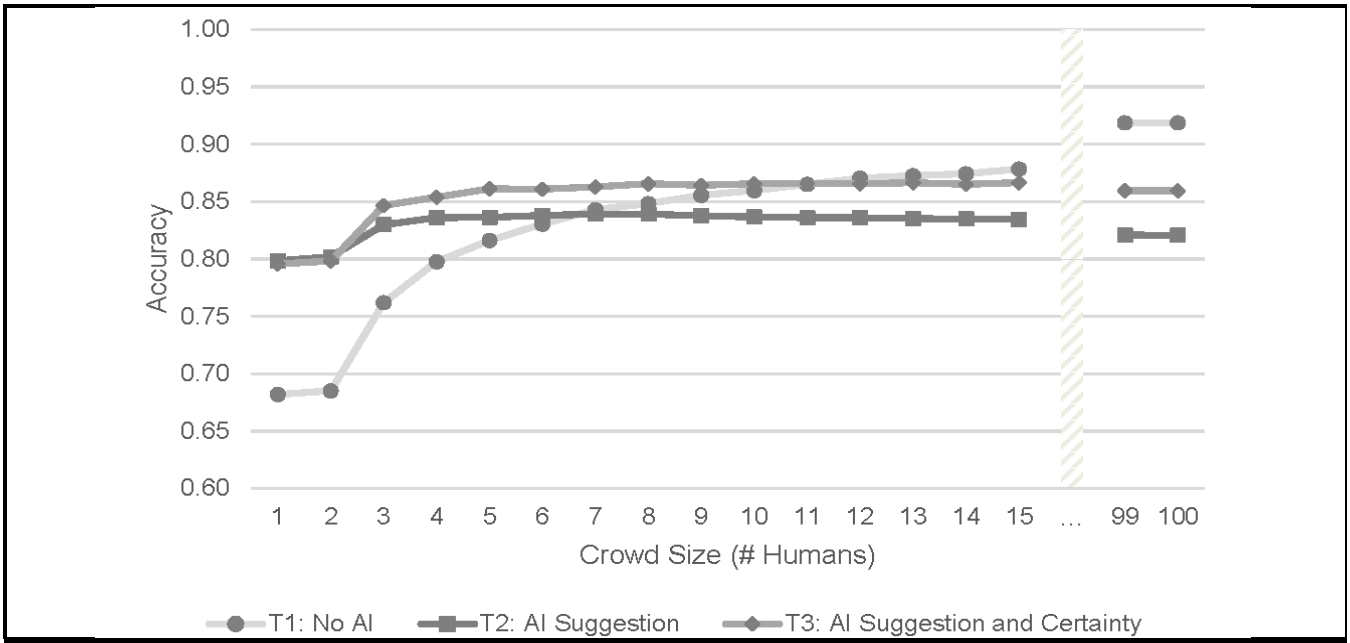

Figure 8. Accuracy of Wisdom of Crowds: Comparison Treatments 1–3 (Experiment 1)

rect answer select the incorrect AI suggestion, but many humans who do not know the correct answer tend to select the AI suggestion, making it the modal choice.

Overall, we can conclude that while AI advice helps individual performance, the reduction in unique human knowledge severely harms crowd performance. When very small in size, pure human groups are outperformed by crowds of humans that work with AI. However, as the group size increases, even at modest sized groups (a computable threshold), the performance of human groups without AI assistance starts dominating performance of those with AI assistance. Finally, interestingly, note that a group of four humans provides more accuracy than the AI.

Wisdom of Crowds: Experiment 3

We now simulate wisdom of crowds based on a second set of 50 images for the three treatments of Experiment 3 with personalized AI advice. Figure 9 plots accuracies for populations without AI advice (Treatment 1'), with permanent AI advice (Treatment 2'), and with personalized AI advice (Treatment 3'). Treatments 1' and Treatments 2' replicate Experiment 1 with fewer images and new subjects. Again, group sizes of one and two lead to the results of single decision makers.

Treatments 1' and 2' behave similar to Experiment 1, suggesting that the subset sufficiently represents the full image set. Similar to Experiment 1, humans in crowds larger than seven members outperform those with AI advice. Interestingly, Treatment 3' outperforms all other treatments for all group sizes. Even for large groups with 100 subjects, humans with personalized AI suggestions are not outperformed by those without AI advice, and outperform those always receiving AI suggestions by more than ten percentage points. This suggests that personalizing AI advice is able to reduce the loss of unique human knowledge while maintaining the benefits of AI advice.

To answer our Hypothesis 4c, we set up a regression analysis as in the previous section, replacing the independent variable AI certainty with AI personalization (1 for Treatment 3'). The regression results are summarized in the right column of Table 6. Again, we obtain a significant negative interaction effect of AI advice and crowd size, finding additional support for Hypothesis 4a. As we find a significant positive interaction effect of AI personalization and crowd size, we find support for Hypothesis 4c. Thus, personalizing the AI's advice mitigates the loss of the crowd benefit of AI advice. Note that both the effect on intercept and the interaction with group size is stronger for personalizing AI advice compared to providing the AI's certainty.

Table 6. Regression Results Group Accuracy in Wisdom of Crowds

	Group Accuracy (Experiment 1)	Group Accuracy (Experiment 3)
Constant (T1)	.726***	.692***
Group Size	.012***	.014***
AI advice	.091***	.109***
AI advice x Group size	-.011***	-.011***
AI certainty	.008*	
AI certainty x Group size	.002***	
AI personalization		0.025***
AI personalization x Group Size		0.004***
Adjusted R^2	.293	.334

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

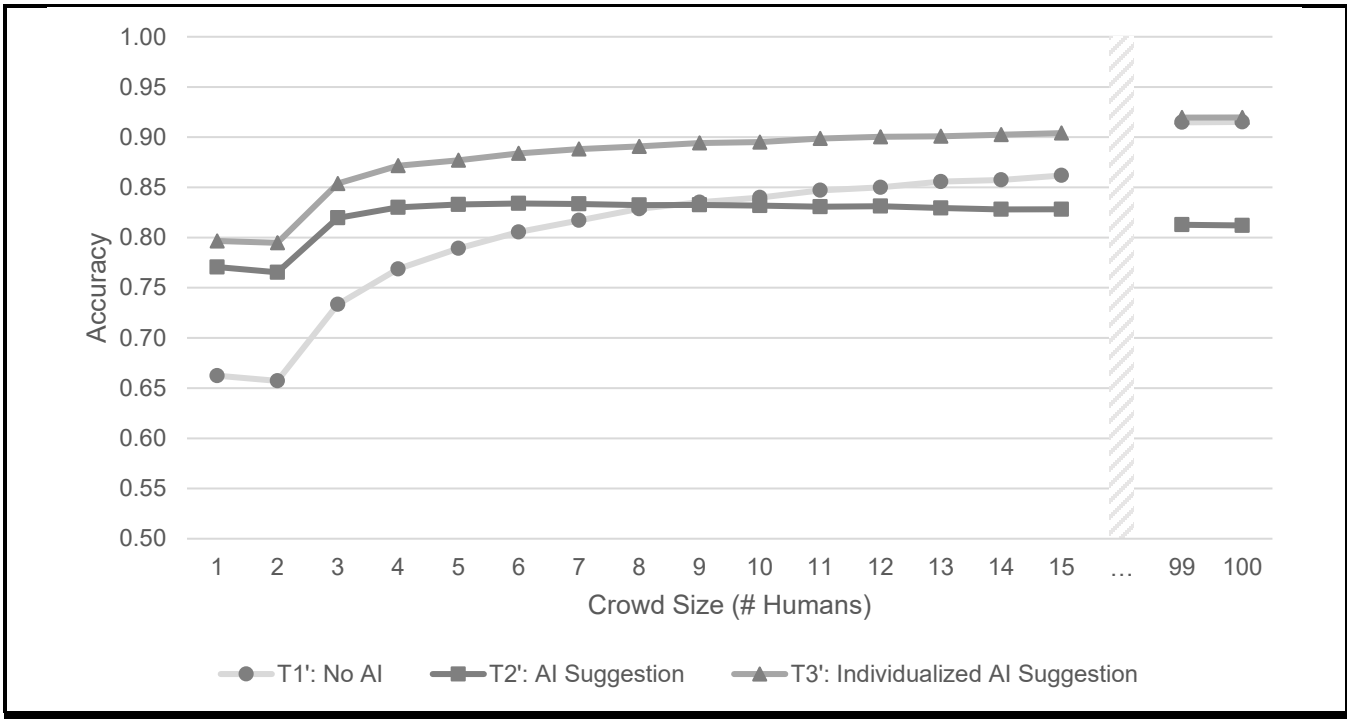

Figure 9. Accuracy Wisdom of Crowds: Comparison Treatments 1'–3' (Experiment 3)

Discussion and Conclusions

Our research suggests that AI advice has two main effects. First, as expected and supported by a host of research, human performance can improve individually by receiving AI advice. However, we also highlight a hitherto undocumented effect: humans lose their unique knowledge, and consequently complementarities with other human decision makers and the AI.

Insights in Human-AI Collaboration

Humans' ability to improve individually from AI advice due to complementary knowledge is well documented in the literature. These complementarities allow AI-advised humans to exceed the AI performance. This potential is not often realized to the full extent, and humans working with AI often perform in between pure human and AI levels (e.g., Zhang et al. 2020). We were able to construct an environment that leveraged AI's ability to assess individual performance to provide selective advice such that our subjects could realize complementarities, surpassing their own and the AI's level of performance. This result is encouraging. We demonstrate that the ability of humans to benefit from AI advice relates to the relative performance and the ability to differentiate between correct and incorrect advice. Thus, humans may even benefit from inferior advice if they are able to pick the instances where the AI performs better.

The second effect, the decrease of unique human knowledge, also relates to complementary knowledge both among humans and between humans and the AI. Each human might follow different decision rules when performing the same task. Our analysis demonstrates that the complementarity reduction due to AI advice has dramatic effects on the wisdom of crowds. Keuschnigg and Ganser (2016) state that for discrete choice tasks, diversity seems to be less important than individual performance. They conclude that for group sizes up to well beyond the size of 20, ability dominates the performance. Our model and our results, however, indicate that groups of humans without AI advice could outperform those with AI advice, even for relatively small group sizes well below the size of 10. Please note that AI is trained by humans: for example, the ImageNet database used in this paper was annotated by MTurk workers. There are reports (Naylor 2021) that human annotators have incentives to behave in line with majority opinions—a behavior that eliminates unique human knowledge and diversity already during the creation of AI algorithms. This further stresses the importance of eliciting unique human knowledge in human–AI collaboration.

Implications for AI-Based Decision Support Systems

Our results have three main implications for the design of AI-assisted decision support environments. First, we demonstrate that humans can realize complementarities with AI using a simple suggestion interface. In many applications where the focus is on the productivity of a single decision maker (or, for example, only two or three decision makers), this might provide a strong feature that can allow firms to combine human and artificial intelligence for superior performance. Second, our results indicate that providing the same AI advice to a whole group of humans has significant downsides. In group decision making situations (e.g., when a group of demand planners is trying to forecast next quarter's sales), complementarities between humans and AI may not be realized although they are crucial. In these situations, AI advice should not be a "one-size-fits-all" solution. Third, modern AI may use its abilities to measure its own certainty and to learn about the human interacting with it in order to overcome this issue. We propose a simple threshold rule that considers the expected harm and benefit of AI advice. It is encouraging that a simple rule like this one works well. Future research should focus on the development of context-specific rules to improve the design of AI-based advice. Overall, our results strongly suggest that application settings have to be considered carefully when deciding if and how AI advice should be built into the decision process.

Limitations and Future Research

We discuss three main areas of future research that that will be fruitful for researchers. First, our results provide a novel direction on the design of decision support environments that include AI assistance to improve human performance both individually and in groups. We demonstrated a way to design personalized AI advice in order to improve both individual and crowd performance. However, there could be other ways in which AI assistance can be combined to benefit from accuracy of AI and diversity of thoughts from humans. Second, we found strong effects of AI advice on the wisdom of crowds and on individual confidence levels. This provides some indication for group behavior with and without AI. These could be challenged by researchers focusing on team behavior and interactions of individuals. The effects in real teams might depend on the type of task, so-called "eureka"

tasks (no specific knowledge is required to solve, and solutions can easily be confirmed when found) typically perform better in real teams (Cooper and Kagel 2005), while this is not necessarily the case for "non-eureka" tasks (Li et al. 2019). Third, one might focus on other forms of interactions with groups of humans or AI. We demonstrated that a group of humans with individual AI advice does perform well. However, there might be alternative setups of interaction that make use of existing complementarities, for example by creating a collaborative environment of AI modules and a group of humans. In general, future studies should investigate whether the loss of unique knowledge affects our capability to work in diverse environments. They should also study the long-term effect of AI advice on human performance in diverse work environments.

Will Humans Become Borgs?

When discussing our research, as in the title of this paper, we occasionally provoke our colleagues by asking whether humans that use AI advice become Borgs. While we intentionally exaggerate, the question triggers a useful debate. Humans improve by using AI, but their behavior converges leading to a loss of their complementarity with the AI and with other humans. This may prevent individuals from thinking about novel ideas and from being more useful contributors to a group. Diversity is a valuable asset which may enable humans to outperform Borgs (and AI) in the long term. However, we also demonstrate that AI could personalize its advice to mitigate this effect, keeping large stakes of human complementarities while benefitting from AI.

References

Angwin, J. Larson, J., Mattu, S., and Kirchner, L. 2016. "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And it's Biased Against Blacks," *ProPublica*, May 23.

Awad, E., Dsousa, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. 2018. "The Moral Machine Experiment," *Nature* (563:7729), pp. 59-64.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. 2019a. "Beyond Accuracy: The Role of Mental Models in Human–AI Team Performance," *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.

Bayati, M., Braverman, M., Gillam, M., Mack, K. M., Ruiz, G., Smith, M. S., and Horvitz, E. 2014. "Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study," *PloS One* (9:10), e109264.

Becker, J., Brackbill, D., and Centola, D. 2017. "Network Dynamics of Social Influence in the Wisdom of Crowds," *Proceedings of the National Academy of Sciences* (114:26).

Bichler, M., Gupta, A., and Ketter, W. 2010. "Designing Smart Markets," *Information Systems Research* (21:4), pp. 688-699.

Bonaccio, S., and Dalal, R. S. 2006. "Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences," *Organizational Behavior and Human Decision Processes* (101:2), pp. 127-151.

Budescu, D. V., and Chen, E. 2015. "Identifying Expertise to Extract the Wisdom of Crowds," *Management Science* (61:2), pp. 267-280.

Carton, S., Mei, Q., and Resnick, P. 2020. "Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity," in *Proceedings of the International AAAI Conference on Web and Social Media* (14), pp. 95-106.

Cohen, J. 2013. *Statistical Power Analysis for the Behavioral Sciences*, New York: Academic Press.

Cooper, D. J., and Kagel, J. H. 2005. "Are Two Heads Better than One? Team Versus Individual Play in Signaling Games," *American Economic Review* (95:3), pp. 477-509.

Da, Z., and Huang, X. 2020. "Harnessing the Wisdom of Crowds," *Management Science* (66:5), pp. 1847-1867.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature* (542), pp. 115-118.

Foster, E. D., and Deardorff, A. 2017. "Open Science Framework (OSF)," *Journal of the Medical Library Association* (105:2).

Galton, F. 1907. "*Vox Populi," Nature* (75:1949), pp. 450-451.

Glikson, E., and Woolley, A. W. 2020. "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals* (14:2).

Hoeffding, W. 1963. "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association* (58), pp. 13-30.

Hong, H., Du. Q., Wang, A. G., Fan, W., and Xu, D. 2006. "Crowd Wisdom: The Impact of Opinion Diversity and Participant Independence on Crowd Performance," in *Proceedings of the 22nd Americas Conference on Information Systems*, San Diego.

Hong, L., and Page, S. E. 2004. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers," *Proceedings of the National Academy of Sciences* (101:46), pp. 16385-16389.

Huber, G. P. 1990. "A Theory of the Effects of Advanced Information Technologies on Organizational Design, Intelligence, and Decision Making," *Academy of Management Review* (15:1), pp. 47-71.

Kelley, E. K., and Tetlock, P. C. 2013. "How Wise Are Crowds? Insights from Retail Orders and Stock Returns," *The Journal of Finance* (68:3), pp. 1229-1265.

Keuschnigg, M., and Ganser, C. 2016. "Crowd Wisdom Relies on Agents' Ability in Small Groups with a Voting Aggregation Rule," *Management Science* (63:3), pp. 818-828.

Kingston, J. K. C. "Artificial Intelligence and Legal Liability," in *Research and Development in Intelligent Systems XXXIII*, M. Bramer and M. Petridis (eds.), Berlin: Springer International Publishing.

Krishnan, H. A., Miller, A., and Judge, W. Q. 1997. "Diversification and Top Management Team Complementarity: Is Performance Improved by Merging Similar or Dissimilar Teams?," *Strategic Management Journal* (18:5), pp. 361-374.

Lai, V., Liu, H., and Tan, C. 2020. "Why Is 'Chicago' Deceptive? Towards Building Model-Driven Tutorials for Humans," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* Honolulu, pp. 1-13.

Lai, V., and Tan, C. 2019. "On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency,* New York: ACM, pp. 29-38.

Leidner, D. E., and Elam, J. J. 1995. "The Impact of Executive Information Systems on Organizational Design, Intelligence, and Decision Making," *Organization Science* (6:6), pp. 645-664.

Li, J., Beil, D. R., and Leider, S. 2019. "Team Decision Making in Operations Management," Working Paper, University of Michigan, Ann Arbor, MI.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. 2011. "How Social Influence Can Undermine the Wisdom of Crowd Effect," *Proceedings of the National Academy of Sciences of the United States of America* (108:22), pp. 9020-9025.

Madsen, M., and Gregor, S. 2000. "Measuring Human–Computer Trust," in *Proceedings of the 11th Australasian Conference on Information Systems*, Brisbane, pp. 6-8.

Moravčík, M., Schmid, M., Burch, N., Lisy, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. 2017. "DeepStack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker," *Science* (356:6337), pp. 508-513.

Naylor, A. 2021. "Underpaid Workers Are Being Forced to Train Biased AI on Mechanical Turk," *Motherboard Tech by Vice*, March 8.

Nijstad, B. A., and Stroebe, W. 2006. "How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups," *Personality and Social Psychology* (10:3), pp. 186-213.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. 2018. "The Preregistration Revolution," *Proceedings of the National Academy of Sciences of the United States of America* (115:11), pp. 2600-2606.

Page, S. E. 2007. "Making the Difference: Applying a Logic of Diversity," *Academy of Management Perspectives* (21:4), pp. 6-20.

Page, S. E. 2008. *The Difference*, Princeton, NJ: Princeton University Press.

Paulus, P. B., and Brown, V. R. 2007. "Toward More Creative and Innovative Group Idea Generation: A Cognitive–Social–Motivational Perspective of Brainstorming," *Social and Personality Psychology Compass* (1:1), pp. 248-265.

Paulus, P. B., Nijstad, B. A., van der Zee, K. I., and Kenworthy, J. B. (eds.). 2019. *The Oxford Handbook of Group Creativity and Innovation*, Oxford, UK: Oxford University Press.

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. 2015. "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, (115:3), pp. 211-252.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* (529), pp. 484-489.

Surowiecki, J. 2004. *The Wisdom of Crowds*: *Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*, New York: Doubleday.

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., and Wojna Z. 2016. "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 2818-2826.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. 2015. "Going Deeper with Convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition,* Boston, pp. 1-9.

Tan, S., Adebayo, J., Inkpen, K., and Kamar, E. 2018. "Investigating Human + Machine Complementarity for Recidivism Predictions," *arXiv preprint arXiv*: 1808.09123.

Wolfers, J., and Zitzewitz, E. 2004. "Prediction Markets," *Journal of Economic Perspectives* (18:2), pp. 107-126.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. 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,* Barcelona, pp. 295-305.

Zhou, J., and Chen, F. 2019. "Towards Trustworthy Human–AI Teaming Under Uncertainty," *IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, Macau, China.

About the Authors

Andreas Fügener is an Associate Professor of Digital Supply Chain Management at the Faculty of Management, Economics, and Social Sciences at the University of Cologne, Germany. His research focuses on the intersection of humans and algorithms, including applications in healthcare. Andreas teaches undergraduate, graduate and executive level courses in general management, analytics, and supply chain management. Prior to joining academia, he served as a management consultant working in the automotive, financial services, and energy industries for several years.

Jörn Grahl conducted this research while at the Faculty of Management, Economics, and Social Sciences at the University of Cologne, Germany, where he served as Professor for Information Systems, Digital Transformation, and Analytics. His research and teaching centers around value creation in the age of digital and AI. He is currently Head of Data Science at the PHOENIX Group,

where he drives the digital transformation of the organization in the areas of data driven decision making and advanced analytics.

Alok Gupta is the Senior Associate Dean for Faculty, Research and Administration at the Carlson School of Management, the University of Minnesota. He is Curtis L. Carlson Schoolwide Chair in Information Management, and the former chair of the Information and Decision Sciences Department. He was awarded the prestigious NSF CAREER Award for his research on dynamic pricing mechanisms on the internet in 2001. He has won numerous awards including ISS Design Science Award twice in 2011 and 2012 and AIS Impact Award in 2020. He was named an INFORMS Information Systems Society (ISS) Distinguished Fellow in 2014 and the Fellow of AIS in 2016. He has served in editorial positions in most major IS journals including senior editor for *Information Systems Research* and *Journal of Management Information Systems*, and AE at *Management Science*. Since 2017, he serves as Editor-in-Chief of *Information Systems Research*.

Wolfgang Ketter is Chaired Professor of Information Systems for Sustainable Society at the Faculty of Management, Economics, and Social Sciences at the University of Cologne, Germany. He is a Director of the Institute of Energy Economics at the University of Cologne, where research he leads focusses on how digital transformation can create a faster and more stable transition to sustainable energy and mobility. He also is professor of Next Generation Information Systems at the Department of Technology and Operations Management, and Director of the Erasmus Centre for Future Energy Business at the Rotterdam School of Management, Erasmus University. He has served on the editorial boards for *Information Systems Research* and *MIS Quarterly*. Since 2017 he is advisor to the German government on energy policy and since 2018 a fellow of the World Economic Forum Global Future Council on Mobility. He won the prestigious AIS Impact Award as well as the INFORMS Wagner Prize Finalists Award in 2020.

Appendix

Mathematical Notation

Individual Decision Making and AI Advice

$\mathcal{T} = \{1..T\}$	Set of tasks
$\mathcal{T}^{AI}$	Set of tasks, where the AI selects the correct option ($a_{t1} = 1$)
$\mathcal{T}^{\overline{AI}}$	Set of tasks, where the AI selects an incorrect option ($a_{t1} \neq 1$)
p_{tc}	Probability, that a human selects option c in task t
p_{tc}^{AI}	Probability, that a human selects option c in task t after receiving AI advice
a_{tc}	1, if AI selects option c in task t, 0 otherwise
e_t	Effect strength of AI advice of task t
e^{AI}	Effect strength of correct AI advice
$e^{\overline{AI}}$	Effect strength of incorrect AI advice
δ_e	Scaling factor effect of correct advice relative to incorrect advice: $\delta_e \frac{e^{\overline{AI}}}{1+e^{\overline{AI}}} = \frac{e^{AI}}{1+e^{AI}}$

Providing AI's Certainty

$p_{tc}^{AI-cert}$	Probability, that a human selects option c in task t after receiving AI advice and the AI's certainty
s_t	Change in the effect of AI advice, e_t, if AI certainty is received, for task t
s^{AI}	Change in the effect of correct AI advice
$s^{\overline{AI}}$	Change in the effect of incorrect AI advice
δ_s	Scaling factor on the effect of receiving the AI's certainty for correct advice relative to incorrect advice with $\delta_s \frac{s^{\overline{AI}}-1}{1+s^{\overline{AI}}.e^{\overline{AI}}} = \frac{s^{AI}-1}{1+s^{AI}.e^{AI}}$

Personalized Suggestions

$\mathcal{H} \in \{1..H\}$	Set of humans
$\mathcal{T}(l_{tc}^{AI} < r_h)$	Set of tasks, where the AI withholds advice for human h
p_{tc}^{AI-per}	Probability, that a human selects option c in task t after receiving personalized AI advice
d_{th}	1, if AI provides advice for task t to human h, 0 otherwise
l_{tc}^{AI}	Ex-ante likelihood that the AI choice c is the correct choice
r_h	Critical ratio of harm of incorrect and the benefit of correct advice of human h